

筑波大学 情報学群 情報メディア創成学類

卒業研究論文

順位推移の視覚的表現を利用した
トレンド分析支援ツールの開発

大久保 心織

指導教員 三末 和男, 志築 文太郎, 田中 二郎

2013年2月

概要

本研究では、時刻と量を伴った順位データの可視化によって、流行の期間や始まりなどのトレンドを分析することを支援するためのツールを作成し、その利用について論じる。このツールでは、順位と量的値を1つの画面に同時に表現している。これにより、従来の表現手法に比べてデータの全体像が把握しやすくなり、個々の事象の流行についてより詳細な分析が可能になった。また、事象の持つテキストラベルから自動的に色相の決定を行うことで、視覚的に事象のカテゴリを確認することができるようになった。

本ツールを用いて、twitter上で話題となっているできごとについてのトピックデータの分析を行った。その結果、ツールを用いることで、トレンドやデータの全体像についての理解が得られた。同時に、メディアに対する人々の反応のデータをツールに適用することで、メディアごとの受け手の反応の差も示せるという可能性も提示した。

目次

第1章	はじめに	1
1.1	時刻付き量的順位データ	1
1.2	時刻付き量的順位データの分析の意義	1
1.3	問題	2
1.4	研究の目的とアプローチ	2
1.5	貢献	2
第2章	関連研究	4
2.1	時刻情報を伴う量的データ推移の可視化	4
2.2	時刻情報を伴う順位推移の可視化	4
2.3	時刻付きテキストデータの可視化	5
第3章	トレンド分析に対する要求	7
3.1	時刻付き量的順位データの概要	7
3.2	トレンド分析に対する要求	8
第4章	表現手法	9
4.1	順位と量の推移の表現	9
4.2	事象の区別	10
第5章	ツールの開発	12
5.1	ツールの概要	12
5.2	基本操作	12
5.3	表現の変更	14
5.3.1	矩形のハイライト	14
5.3.2	矩形の整列順序	15
第6章	ユースケース	17
6.1	データの概要	17
6.2	観察	20
6.2.1	流行の全体像の把握	20
6.2.2	量的値の変動の把握	21
6.2.3	トピックのカテゴリの把握	21

6.2.4	データの全体像の把握	21
6.3	考察	24
6.4	ツールの有用性に関する議論	26
6.4.1	事象の連続性について	26
6.4.2	データの傾向把握について	26
6.4.3	テキストラベルからの色相自動算出について	26
第7章	おわりに	28
7.1	まとめ	28
7.2	今後の課題	28
	謝辞	30
	参考文献	31

図目次

4.1	表現の概要	9
4.2	量的値の表現	10
5.1	ツールの概要	13
5.2	マウスカーソルを矩形に乗せた場合の表示	13
5.3	テキストラベルの一覧表示	14
5.4	事象の詳細表示	14
5.5	順位の表現	15
5.6	事象の連続性の表現	15
5.7	カテゴリ順（色相順）での配置	16
6.1	twitter トピックデータを適応させたツールの外観	19
6.2	単発的なイベントの発生	21
6.3	『天空の城ラピュタ』に関するトピックの発見と観察	22
6.4	サッカー女子日本代表の W 杯優勝に関するトピックの発見と観察	23
6.5	1 月 1 日（左端）の矩形とその他の軸の矩形の比較	24
6.6	カテゴリの把握	25

表 目 次

1.1 時刻付き量的順位データの例（商品の売上）	1
6.1 twitter トピックデータの例	18

第1章 はじめに

1.1 時刻付き量的順位データ

世界にはさまざまなランキングが存在する．例えば，学校のテストの成績や，商品の売上，あるいは流行歌などがランキングを伴って表されているものは我々にも馴染み深い．このランキングを算出する際に特に重要になるのが，その事象が持つ量であり，商品ならばいくつ，あるいはいくら売れたか，検索キーワードならばそれが検索された回数という量をもって，ランキングが決定されることもしばしばある．

本研究では，ある事象に対して出現回数や生成量などの量（以下量的値），それを利用して付けられたランキング（以下順位），及び単位時間ごとに計測された時点の時刻情報が定義されたものを「時刻付き量的順位データ」とする．例として表 1.1 に示されるようなデータが挙げられる．

表 1.1: 時刻付き量的順位データの例（商品の売上）

時刻	順位	事象	量的値
2013 年 1 月 1 日	1	日本酒	100
2013 年 1 月 1 日	2	コーラ	90
2013 年 1 月 1 日	...	...	...
2013 年 1 月 2 日	1	オレンジジュース	120
2013 年 1 月 2 日	2	緑茶	85
...	...	...	...

1.2 時刻付き量的順位データの分析の意義

時刻付き量的順位データを分析することは，マーケティング，政治情勢分析など，流行を常に追いつけるさまざまな分野において非常に重要である．その中でも，時間によって変動する順位を見ると，いつ・どのようなものが流行したかという観点から，当時の世界のようすを知ることができる．また，過去の順位を参考にして，これからどういうものが流行するのかという予測をすることもできる．

1.3 問題

現代の情報社会化，人々の興味関心の分散などから，社会の中にある時刻付き量的順位データの量は増加する一方である．したがって，ただそれらの生データを見るだけでは分析を効率良く行うことができない．加えて，流行を知るには順位の変遷等の全体像を把握する必要がある．そのために，データを視覚的に表現する手法（可視化）が広く用いられている．

我々にも馴染み深い表現方法として，表 1.1 のような表に，「前の時刻との順位と比較」という列を付け加え，前の時刻に比べて順位や量的値がどれだけ増えたか，あるいは減ったかを表す表現方法がある．増減に対応して上向き・下向きの矢印も付随しているものもある [1, 2]．しかし，これをある時間帯に対して並べた場合，流行の全体像というものは読み取りづらい．

ひとつの事象を 1 本の線に表し，横軸に時刻を，縦軸に順位をとったような折れ線グラフを用いて視覚的な表現を得ようとすることも考えられる．だが，これでは同時に事象の持つ量的値を表現することができない．かつ，大量のデータを表現しようとすると，その事象の数に応じた本数の線が必要になり，視覚的に非常に混雑してしまう．

1.4 研究の目的とアプローチ

本研究の目的は，時刻付き量的順位データをもとにトレンドの発見や分析を支援することである．これに対するアプローチとして，ひとつの画面で一定期間についての事象の順位と量的値の推移が見えるようにデータを視覚的に表現することを設定する．また，その表現を利用したトレンド分析支援ツールを作成する．

本研究では，マーケティング分野などでユーザが分析の対象としたい任意の時刻付き量的順位データについて，どのような事象が流行しているか，さらにはその事象がいつ流行り始めたか，いつ流行が終息したか等の情報を得られるような可視化を提供するツールを作成することを目標とする．

1.5 貢献

本研究の貢献として挙げられるのは以下の 2 点である．

1. 順位と量的値を同時に表現することで，データの全体像が把握しやすい表現手法を開発した
2. 事象の持つ文字列を利用して，事象のカテゴリを反映した色相値の自動算出技術を開発した

1 では，より詳細なトレンド分析が行えるように，従来の表現手法では一度に表すことが難しかった順位と量的値を同時に表現する手法を提示した．両方の変量の推移を観察することで，片方のみの観察では得られなかった特徴も見えるようになると期待される．

2について、大量のテキストデータを文字のまま画面に表そうとすると、視覚的に混雑しとても分かりづらくなる。加えて、ひと目で眺めただけでは各事象を区別することや、どの事象が似たもの同士かを判断することも困難である。そこで、事象に付随するテキストラベルの文字列を利用し、ハッシュ法をもとにした色相の割り当てを行った。これによって文面が似たものに近い色相が割り当てられるようになり、事象を色相で見分けることができるようになった。

第2章 関連研究

本研究では、時刻情報と量、および順位を持つデータに対して、それらを一度にひとつの画面に表すという表現手法を用いる。このような、時刻情報を伴う量的データの推移、あるいは順位の推移の可視化はこれまでも数多く行われてきている [3].

さらに、事象を指し示す文字列に着目すると、ここでの可視化はテキストデータの可視化とも言うことができる。そこで、同様に時刻付きテキストデータの可視化に関する関連研究についても記述していく。

2.1 時刻情報を伴う量的データの推移の可視化

時刻付き量的データの可視化で最も一般的なものは線グラフ (Line Chart) であり [4], 量をよりの確に捉えられるように改良したものが積み上げ面グラフ (Stacked Area Chart) である。積み上げ面グラフを利用した時刻付き量的データの可視化は Wattenberg の NameVoyager[5] や Lee の The GTD Explorer[6] があり, これらはインタラクティブ操作でユーザの見たい情報を絞込みながら積み上げ面グラフで可視化を行なっているものである。

積み上げ面グラフを発展させた Harve らの ThemeRiver[7] もまた, よく用いられる手法である。これは「川 (River)」のモチーフを用いることで, 時系列に沿った量変化を直観的に見られるようにしたものである。積み上げた図形を上下対称に配置することで, 推移の様子が形の変化として把握しやすくなっている。

このような, 図形の幅すなわち長さで量を表現することは, 人間の認知的にも理にかなったものである [8]. しかし積み上げ面グラフや ThemeRiver のような各事象を層として重ねて表示するものは, 時刻に沿って層を入れ替えていくことができない。つまり, ある時点における重なるの順番と, 別の時点の重なるの順番を入れ替えることが難しい。本研究で扱う時刻付き量的順位データにおいて, 順位を積み上げられた順番で表現しようとした場合, そうした入れ替えが頻繁に発生し, 提示することが困難になる。

2.2 時刻情報を伴う順位推移の可視化

第1章で述べたように, 実世界に存在する順位データの多くは時刻付きデータであり, 我々の生活の中にもこうした時刻付き順位データ, およびその可視化は広く浸透している。例として, Yahoo!Japan の検索ランキング [1] では, その日に検索されたキーワードでランキング

表を作成しており、順位が前日と比べて上がったか・下がったか・そのままか、あるいはランキング圏外から登場したのかということが矢印などの記号で分かるようになっている。

また、The Economist の 2009 年 9 月 11 日付の記事 [9] で示されている可視化は、2009 年から 2010 年にかけて各国でのビジネスの行いやすさの格付けがどう変わったかということを示している。ここでは横方向に順位が割り当てられており、左にいくほど順位が高いということを表す。2009 年と 2010 年での順位をそれぞれプロットし、2009 年からどれだけ順位が変化したかを右向きまたは左向きの矢印で表現することで、順位がどれだけ上昇または下降したかが分かる。この場合、すべての矢印が左を向いているので、ここに示された国の順位はすべて上昇しているということが見て取れる。

折れ線グラフを拡張した順位表現として、Brinton の Rank Charts[10] が挙げられ、我々のよく目にするところでは駅伝の現在までの順位表示などに利用されている。さらに、この応用として Slopegraph[11] や Bumps Chart[12] も似たような表現を用いている。

Brinton は Rank Charts をさらに発展させ、その時点の量的データをラベルの上に表現する手法も提案している [10]。このような表現は The Telegraph の 2011 年 8 月 13 日付の記事 [13] にも見受けられる。本研究の表現はこれらに基づいたものであるが、より多くのデータを観察するため、事象を自動で色分けしてカテゴリ化するなどの工夫や、前の時間帯からの順位の移動がよりわかりやすくなるような表現を付加している。

順位と量を同時に表現したものとして、ThemeRiver を拡張した Shi らの RankExplorer[14] が存在する。これは事象となる検索キーワードを順位でカテゴリ分けしており、順位の推移とともに各カテゴリへの流入などを棒グラフを埋め込んで表したものである。しかし、ThemeRiver や積み上げ面グラフのような表現で順位を表すことは、層の入れ替わりなどを考慮しなければならず困難である。RankExplorer は順位でカテゴリ分けすることで、層の入れ替えを発生させないようにしているが、本研究では個々の事象の順位の推移がより詳細に見られるような可視化手法を提案する。

2.3 時刻付きテキストデータの可視化

ThemeRiver や、その表現に近い「流れ」を利用したテキストデータの可視化として Liu らの TIARA[15, 16] や Cui らの TextFlow[17] が存在するが、前述の通り層を途中で入れ替えるのが困難であり、順位を伴った表現には向かない。

折れ線グラフに準拠する Tufte の Sparkline[18] をタグクラウドに埋め込んで、単語のトレンドを表現するのが Lee らの SparkClouds[19] である。これはどの事象が関連性を持っているかなどの比較をすることは難しい表現である。

同様にタグクラウドを利用したものとしては Collins らの Parallel Tag Cloud[20] が挙げられる。これはカテゴリに分けられたタグクラウドが並行軸状に配置されており、このカテゴリ分けを時間で行えば、時系列テキストデータの表現も可能になる。この表現は単語が順位順ではなくアルファベット順に並べられている。また、実際の文字列を画面上に表示しており、量によって文字の大きさが変化しているため、順位順に単語を並べると、大きい文字の単語

がまとまったところに配置され，視覚的混雑が増して見つらくなるという欠点がある．重ならないよう十分スペースをとるとしても，小さい文字ばかりが存在するスペースがより無駄になってしまう．順位付きの大量の時系列テキストデータを表示する際に，領域の確保が難しくなるのである．

第3章 トレンド分析に対する要求

3.1 時刻付き量的順位データの概要

本研究で扱うデータは「事象が出現回数等の量的値により順位付けされた時刻付きデータ」である。具体的に言うと、例えば商品の販売データであれば、事象はある商品が購入されたこと、量的値は購入された数や売上金額となり、事象を区別するためのテキストラベルは商品名から付けられる。また、政党支持率であれば、事象は政党に対する支持、量的値は実際の支持率であり、テキストラベルは政党名から付けられる。

このような量的値を元に順位付けを行い、データを採取した時刻を付加したものを実際のデータとする。この時、データは単位時間ごとに計測され、例えば単位時間を1日、あるいは1時間と設定した場合、1日おき、あるいは1時間おきにデータが計測される。

数式による定義 事象の集合を A で表す。すなわち、

$$A = \{a_1, a_2, \dots, a_m\}. \quad (3.1)$$

また、時刻の集合を T で表す。

$$T = \{t_1, t_2, \dots, t_n\}. \quad (3.2)$$

時刻 t における事象 a のもつ量を $f(a, t)$ で表す。ここで f は次のような関数と考えることができる。

$$f: A \times T \rightarrow \mathbb{R} \quad (3.3)$$

同様に順位を $r(a, t)$ で表す。 r は次のような関数と考えることができる。

$$r: A \times T \rightarrow \{1, 2, \dots, m\} \quad (3.4)$$

ある時刻 t において同順位のものはないとする。つまり、 $a \neq b (a, b \in A)$ であれば $r(a, t) \neq r(b, t)$ とする。

順位 $r(a, t)$ は f にもとづいて決められた $g(a, t)$ の大きいものから順に1位、2位…と定められる。例えば $g = f$ と定義することができるが、他の定義の仕方も存在する。

3.2 トренд分析に対する要求

広義の「トレンド分析」とは、過去の一定の期間分のデータを観察し、さまざまなデータと照らしあわせて事象の動向を読み取るものである。得られた結果を利用して今後の情勢の展開を予測することもトレンド分析の重要な役割である。すなわち、「過去」のできごとから「現在」の状態を把握し、「未来」についての見通しを立てるといった、時間の流れに対して全方面的な観察が要求される。より具体的に言えば、ある事象についてのトレンドを観測するとき、どのようなものが流行したか、またはしそうか、流行した（している）ものがいつ頃注目され始めたか、ということ进行分析する。

本研究で扱うデータによるトレンド分析も、このような分析行動に則して定義することができる。したがって、本研究におけるトレンド分析に対する要求を以下のように設定する。

要求 1 何が流行しているか・していないかを知りたい

要求 2 いつごろ流行し始めたか・流行が終了したかを知りたい

このような要求が過去・現在・未来に対して生じるということである。

すなわち、過去どのようなものが流行していたかと振り返ること、今どのようなものが流行しているかを観察すること、これからどのようなものが流行しそうかと予測することをユーザは望む。なお、ここで想定されるユーザとはマーケティング分野などにおいてトレンド分析を行う人である。

第4章 表現手法

この章では、第3章で述べた要求を満たすような視覚的表現の設計について述べる。これは、トレンドの動向や全体像の把握ができるよう、事象の持つ順位と量的値を同時に、時間に沿って表現する手法である。

4.1 順位と量の推移の表現

本研究で提案する可視化は Inselberg らの Parallel Coordinates[21] や Kosara らの Parallel Sets[22] に発想を得た構造になっており、ここでは1本の並行軸が1単位時間を表している。すなわち、縦に並ぶものは同じ時刻情報を持つ事象であるということになる。

1つの事象は1つの矩形として描かれ、その時刻に発生した事象が垂直に並べられている(図4.1)。事象の持つ順位が高いほど、矩形は上の方に描画される。次の時刻で計測されたデータにおいても同じ事象が観測されていた場合、その矩形同士を太い線(リボン)でつなぐ。反対に、連続していなければリボンを描画されない。これによって、事象の連続性や断続性が把握しやすくなる。

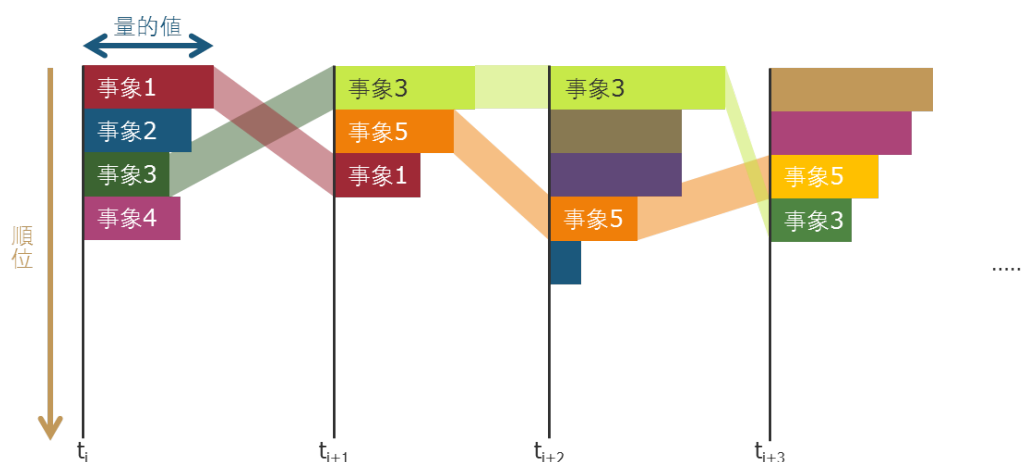


図 4.1: 表現の概要

順位の表現 順位 $r(a, t)$ を表現する要素として縦方向の座標を用いる．各軸で上にあるものほど、その時刻において高い順位であることを表し、上から順に 1 位, 2 位... と並ぶ．順位が高いものを物理的に高い位置に置くことで、より直観的な表現を提示する．

量的値の表現 矩形の横幅が量的値 $f(a, t)$ に対応している．

例えば HistoryFlow[23] では、縦幅で量的値を表現しており、これによって全体の量の推移も一目で分かるようになっている．一方、本研究のように縦の並びで順位を同時に表そうとしたとき、横に並ぶもの同士では順位が必ずしも一致しない．これは縦幅を固定していないために生じる現象であるが、本研究ではこの事態を避けるために、矩形の横幅で量的値を表すことにした．

ただし、量変化を観察したい場合もあるということを考慮し、矩形の縦幅が量を表現するように表示を切り替えることができるようにもする．これによって HistoryFlow の表現により近くなり、各事象の量だけでなく、その時刻における事象の総量が軸の長さで表現されることになる．

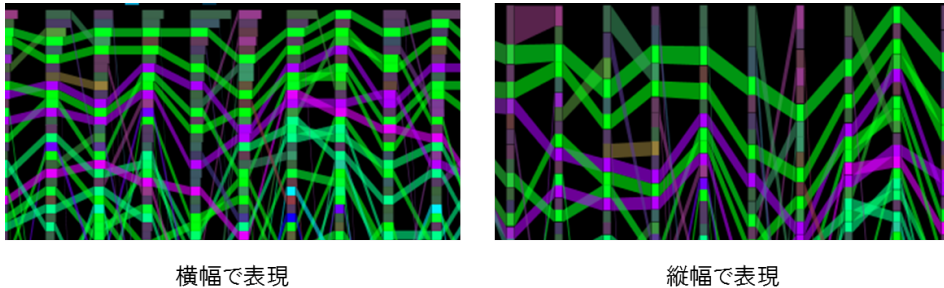


図 4.2: 量的値の表現

時刻情報（時間軸）の表現 時刻 T を表現するのは横方向の座標である．画面の左側にある軸ほど古い時刻であることを表しており、左から右へと古い順に並んでいる．ある時刻における表現を並列させることで、時間の推移に伴う事象の変動を観察することができる．

4.2 事象の区別

事象である A の要素 a をそれぞれ区別して表示するため、矩形を塗りつぶす色相を利用する．矩形を塗りつぶす色相については、その矩形が表す事象に予め付けられたテキストラベルによって決定される．各事象は、それ自体を他の事象と区別できるように、それぞれ一意なテキストラベルを持っている．よって、視覚的表現においても、事象を区別する色相の算出にテキストラベルを用いることは有意である．また、この色は、次の時刻において順位が上昇していれば彩度と明度が上げられ、より目立つようにしている．

事象に付けられたテキストラベルから色相を算出する方法として、以下のような計算を行う．

1. テキストラベルとなっている文字列の各文字（文字コード）を数値に変換する.

長さ n の文字列 s を $s = \langle c_1, c_2, \dots, c_n \rangle$ と定義する.

また, 文字 c の数値で表された文字コードを $D(c)$ とおく.

2. それらをすべて足しあわせたものを文字長で割り, 1 文字あたりの平均値を求める.

文字列 s の数値は以下によって表される.

$$D(s) = \frac{1}{n} \sum_{i=1}^n D(c_i) \quad (4.1)$$

3. 求めた値を色相の最大値で割り, 出た余りを色相値とする.

色相の最大値を h とおけば, 求める色相値 $H(s)$ は

$$H(s) = D(s) \bmod h \quad (4.2)$$

この手法は Wattenberg らの Chromograms[24] を参考にしている. 彼らは単語の頭 3 文字を利用して色相, 彩度, 明度の 3 値を定義しているが, 本研究では彩度および明度は順位の上昇を表すことに利用しているため, 上記のようになるべく色相のみで文字列を一意的な値で表現できるようにした.

数値の平均化 提案する手法による色相の割り当てでは, テキストの数値に対して平均化を行なっている. これは, 同じ事象に対する表記のバリエーションや, 似た事象に対する似た言葉を含む文字列について, 差異を減らしなるべく近い色相値を算出するために行うものである.

テキストラベルは事象を区別するために付けられているが, 同じような事象に対してはしばしば同じようなテキストラベルが付けられる. 例えば, 商品の売上データを考えたとき, 「ミルクチョコレート」や「ビターチョコレート」, 「アーモンドチョコレート」などは, テキストである商品名こそ違うが, すべて「チョコレート」というカテゴリに属する商品であると言える. これらを表現する矩形に対して似たような色相を割り当てるため, 平均化を行う.

また, カテゴリごとに事象が把握できることで, カテゴリ全体の流行の推移も見えるようになる. 先の商品の売上データで言えば, 2 月 14 日のバレンタインデーにあわせ, その付近において「チョコレート」というカテゴリの商品の売上が伸びていることが確認できると期待される.

剰余の利用 剰余を利用することで, 文字列の数値がどのような最大値をとるか分からない場合においても, 色相の最大値以下の数値を割り当てることができる. これは剰余を用いたハッシュ法である.

第5章 ツールの開発

5.1 ツールの概要

本研究では時系列量的順位データを可視化し，トレンド分析を支援するツールを開発した．これは Shneiderman の Visual Information Seeking Mantra[25] の原則にもとづき，はじめにデータの概要を俯瞰し，それから各事象の詳細を観察できるような構造とした．

4章で述べたデータの視覚的表現は，ある期間における事象の順位と量的値の推移を1つの画面に表示し，トレンドの全体的な流れを提示するものである．ここからさらに各事象について詳細な分析が行えるように，インタラクティブ操作を利用しながらトレンド分析を行うツールを作成する．

本ツールは Processing¹によって開発を行い，GUI のライブラリとして ControlP5²を利用している．

5.2 基本操作

ツールの画面は大きく3つに分けられる (図 5.1) ．

視覚的表現 画面の中央には，第4章で設計したメインの視覚的表現が表示されている (図 5.1 の 1) ．この図では，時刻の集合 T の個数 n は 31，事象の集合 A の個数 m は 100 である．

ひとつの事象を表す矩形にマウスカーソルを乗せると，事象の量に応じた大きさの円および事象のテキストラベルすべてが表示される (図 5.2)．

テキストラベルの一覧表示 画面中央において，ある時刻をキーボードの矢印キーの左右キーで選択すると，その時刻のデータ中にある事象の順位とテキストラベルの一部が画面右部に一覧表示される (図 5.1 の 2 および図 5.3) ．なお，文字の大きさは事象の量に対応しており，文字列の左端には事象に割り当てられた色相の円と順位を表示している．

事象の詳細表示 画面中央において，矩形をクリックすると，その矩形が表現する事象の詳細が画面下部に一覧表示される (図 5.1 の 3 および図 5.4) ．画面に描画しきれない分については，右側のスクロールバーの三角形をクリックすることで見ることができる．

¹<http://processing.org/>

²<http://www.sojamo.de/libraries/controlP5/>

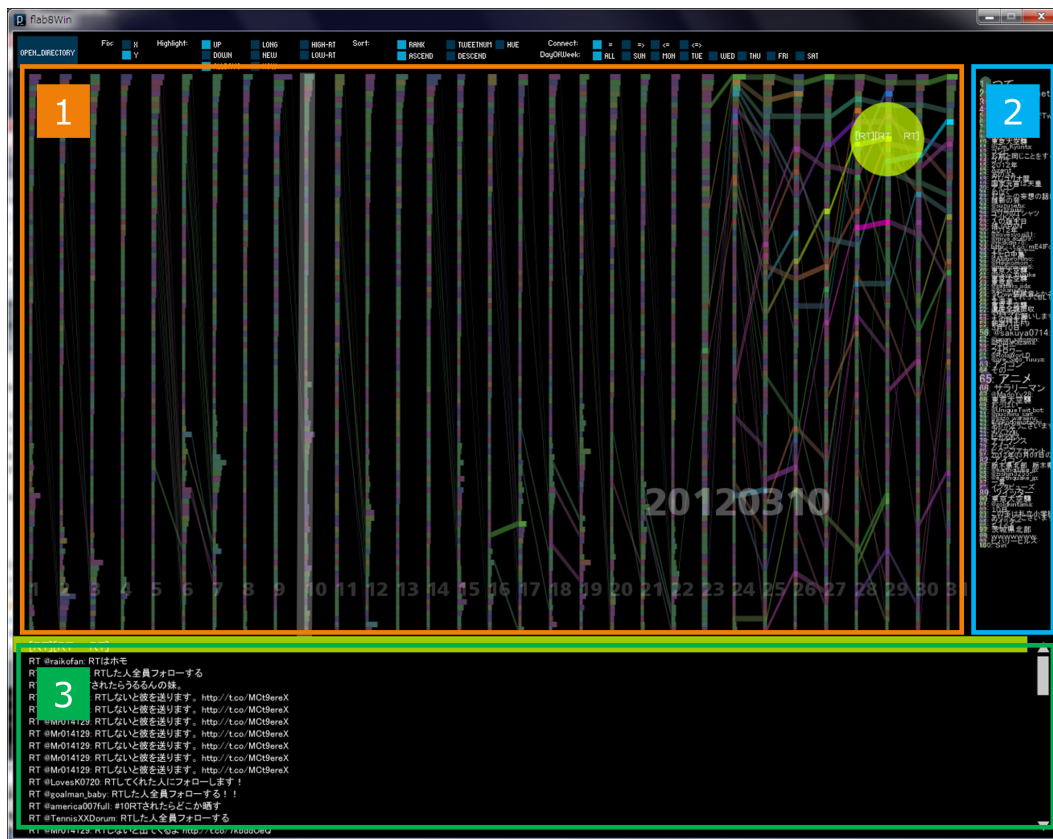


図 5.1: ツールの概要



図 5.2: マウスカーソルを矩形に乗せた場合の表示

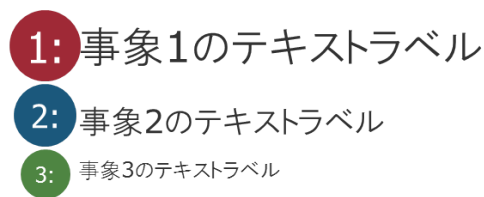


図 5.3: テキストラベルの一覧表示



図 5.4: 事象の詳細表示

5.3 表現の変更

各事象が持つ属性の値に応じて、矩形に対して強調したいものを変更することができる。本ツールにおいて、強調することに利用する視覚的要素としては、矩形の幅と明度・彩度がある。特に、明度と彩度で強調する場合を「ハイライト表示する」と呼ぶ。また、同様に、上下の配置を決定する属性を変更することもできる。

5.3.1 矩形のハイライト

ある属性について、どのような値や特徴を持つものをハイライト表示するかをひとつ選択できる。これによって、特定の事象を強調することができ、目的の動向を示しているような事象をより観察しやすくなる。

順位の上昇と下降 デフォルトでは順位が上昇している事象をハイライトするように設定した。流行を探索するにあたり、もっとも参考になるのは順位の上昇であると考えたためである。

反対に、順位が下降しているものをハイライトすることも可能である。この強調によって、流行が終息しつつあるような事象を確認することができる。

事象の連続と単発 順位の上昇・下降に関係なく、連続して出現しているものをハイライトすることができる。同様に、他の時刻では全く出現しなかったような、単発的な事象もハイライト表示ができる。この表示によって、長期にわたり人々に支持された事象や、ある時刻においてのみ人々が興味を持った事象などが分かる。

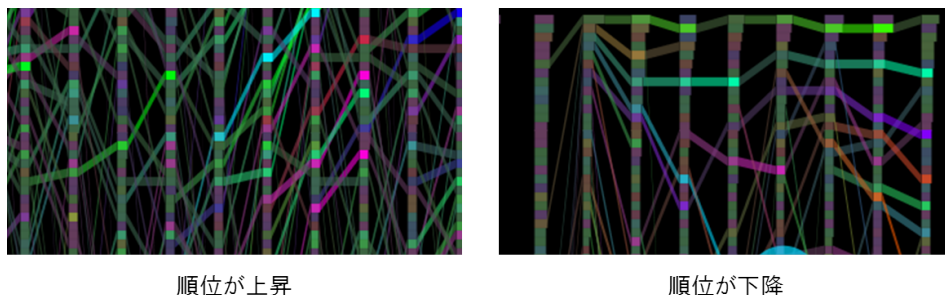


図 5.5: 順位の表現

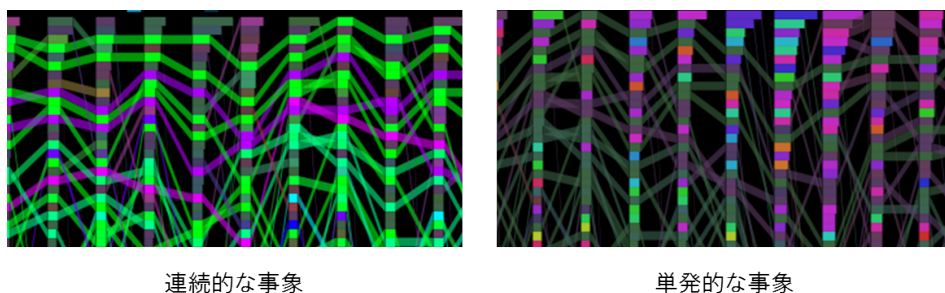


図 5.6: 事象の連続性の表現

事象の属性の大小 事象が順位と量的値以外に何らかの属性を持っており、それが二値であったり、あるいは量的なものであった場合、値にしたがってハイライト表示する。

例として、商品の購買データで、その商品が期間限定商品であるかないかという二値属性を持っていたとする。このとき、「期間限定商品である」ものをハイライトすることもできるし、「期間限定商品でない」ものもハイライトすることができる。

これにより、その事象の持つ順位と量的値以外の属性に着目した分析を行うことができる。

5.3.2 矩形の整列順序

各事象を表す矩形の上下の配置を決定する。整列はすべてにおいて昇順か・降順かを選択できる。

順位順 デフォルトでは、事象の順位が高いものほど、それに対応する矩形が画面上部に配置される。本研究で用いるデータは、様々な属性がついているが基本的には「順位データ」であり、これを観察するためにデフォルトではこの設定にしている。

量的値順 事象の持つ量的値にしたがって上下の配置を決定できる。量的値に着目したいときに用いると、値の多いもの、あるいは小さいものが把握しやすくなる。

カテゴリ順 事象に割り当てられた色相にしたがって上下の配置を行う。文字表現として近いものが似た色相を割り当てられるため、ある程度似たカテゴリに属するテキストラベルを持つ事象が近くに表示される。すなわち、対象のデータにおける事象はどのようなカテゴリに分けられるかが分かる。加えて、カテゴリごとの事象の数の差も、異なる色相の矩形の数によって分かる。

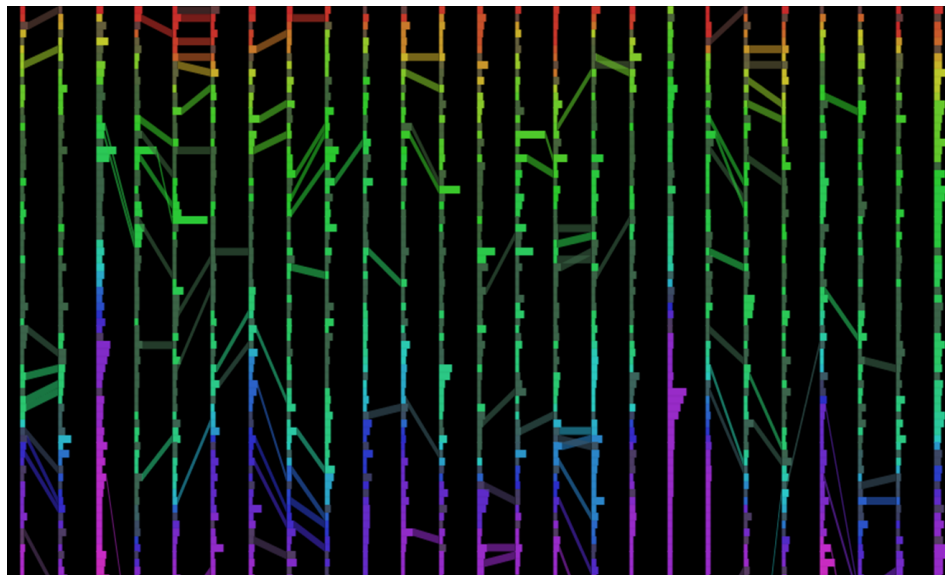


図 5.7: カテゴリ順（色相順）での配置

第6章 ユースケース

実際のデータをツールに読み込ませ、表現を変化させることで観察できた具体的なトレンド分析の例を示す。データはソーシャルメディアにおいて話題になっている単語群を順位付けしたものであり、ここから人々がどのような話題に関心を抱いているかといったことを分析していく。トレンド分析への本ツールの利用法を提示し、分析で得られた結果について考察したのち、それをもとにツールの有用性について議論を行うことを目的とする。

6.1 データの概要

このユースケースで用いるデータは短文投稿サービス **twitter**¹ に投稿された文書（ツイート）を整形したトピックデータである²。

ある1日のツイートに含まれる共通の単語群を「トピック」と定義する。そのトピックの順位、トピックを構成する単語群、関係する実際のツイート数、および実際のツイートが1つのレコードとなる。これが1日ごとにまとめられており、2011年7月～2012年6月の1年間分のデータが用意されている。このデータにおいて、量的値はツイート数、テキストラベルはトピックとなる単語群である。なお、「実際のツイート」に含まれる属性は **twitter** のユーザ ID、ツイート ID、本文、時刻情報、GPS 情報であるが、今回は本文のみを画面上に表示することにした。

データの属性として、順位・量的値・時刻の他に「リツイート (RT)³率」という属性を加える。これは、あるトピックに関する実際のツイートのうち、RT であるものがどれだけの割合を占めているかということを計算したものである。これをもとにして、RT 率が高いもの・低いものをハイライトできるようにした。

テキストラベルからの色相算出については、トピックとなっている単語群を頭文字の文字コード順にソートし、その1単語目となった単語を利用した。これは、各トピックについて、同じ事柄を話題にしているものは、なるべく同じ単語を用いて色相の算出が行えるように配慮したものである。

データの具体例を表 6.1 に、実行例を図 6.1 に示す。

¹<http://twitter.com/>

²株式会社富士通研究所から提供を受けた。トピック判別方法並びに順位算出方法 2 についても提供を受けたものである。

³そのツイートが他人のツイートを参照したものであることを表す

表 6.1: twitter トピックデータの例

時刻	順位	トピック	ツイート数	実ツイート (本文のみ)
2011 年 7 月 1 日	1	七夕の短冊, 一	236	【七夕の短冊】一 リア 充 になりたい hand.ko] http://t.co/NWUEaFU
				【七夕の短冊】一 とにか くお金が欲しい 角都] http://t.co/ZMakn9f
				...
2011 年 7 月 1 日	2	コスプレ, もののけ姫, もののけ姫のコスプレ, http://twitpic.com/5jirad, @supantutarou:	84	RT @supantutarou: もの のけ姫のコスプレ探して たんだが, このコスプレ は予想外だった... #ntv http://twitpic.com/5jirad
				RT @supantutarou: もの のけ姫のコスプレ探して たんだが, このコスプレ は予想外だった... #ntv http://twitpic.com/5jirad
				...
2011 年 7 月 1 日	...	...	...	...

このトピックデータは株式会社富士通研究所から提供を受けたものである。トピックの判別及び順位の算出方法は以下の通りである。

トピックの判別 処理の対象となるツイートの集合（1 日分など）において、指定回数以上出現したキーワードを抽出してトピックとする。このキーワードとは 2 つ以上の単語からなる組み合わせであり、単語とは形態素解析によって名詞と判断されたもの、オンライン百科事典 Wikipedia⁴に登録されている 3 文字以上のエントリー、URL、ハッシュタグ⁵である。

順位の算出 本ユースケースで用いたデータにおける順位の算出方法には 2 種類ある。算出方法 1 はツイート件数のみに基づいた順位付け、算出方法 2 はトピックの新規性を重視した順位付けである。

それぞれの算出方法において、トピック a の時刻 t における順位 $r(a, t)$ を決定する関数 $g(a, t)$ は以下で定義される。

⁴<http://ja.wikipedia.org/wiki/>

⁵twitter において、投稿したツイートに対して投稿者が付けるタグ

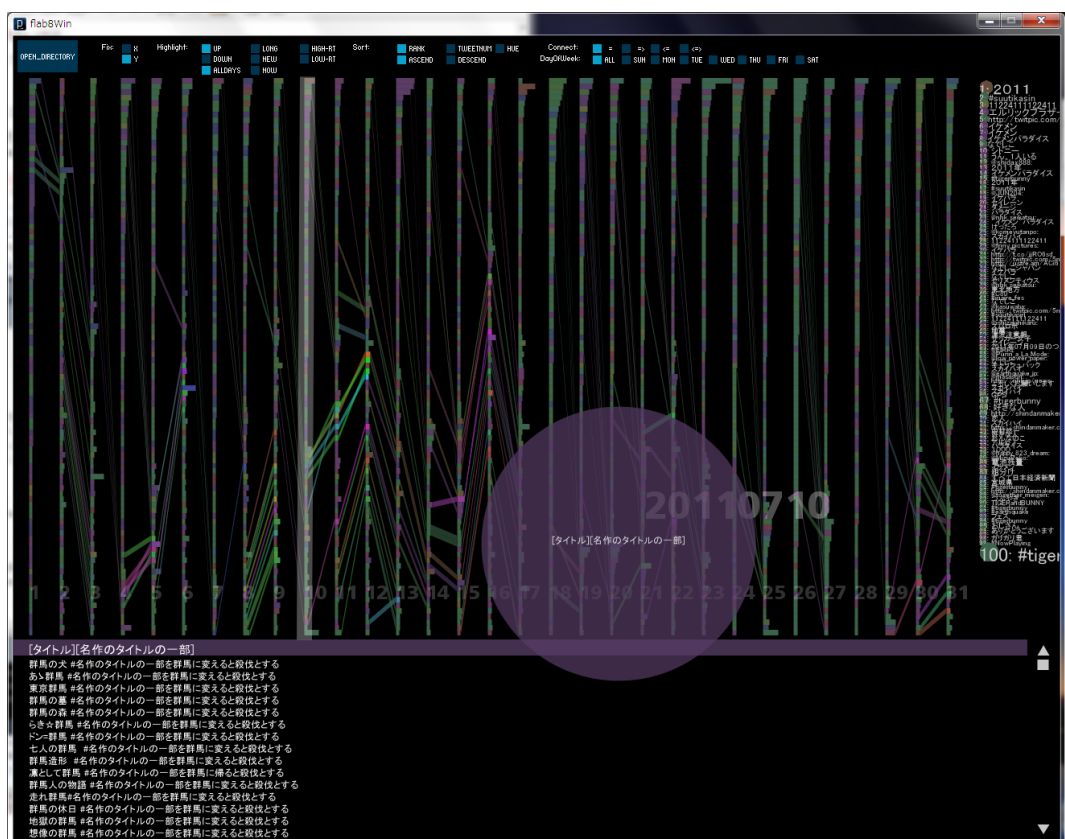


図 6.1: twitter トピックデータを適応させたツールの外観

算出方法 1

$$g(a, t) = f(a, t). \quad (6.1)$$

ここで $f(a, t)$ は時刻 t におけるトピック a の持つツイート数である.

この算出方法は、順位を求めるにあたり最も基本的な方法である.

算出方法 2

$$g(a, t) = \frac{fq(a)}{all(a) + 1}. \quad (6.2)$$

ここで $fq(a)$ はトピック a の出現頻度で、時刻 t におけるツイートの集合のうち、トピック a となるキーワードを含むツイートの数である.

また、 $all(a)$ は過去にトピック a となるキーワードがトピックとして選択された回数である. 今回のデータでは、2010 年のツイート集合からトピックの判別を行った結果をもとに計算している.

この算出方法はトピックの新規性が高いと順位が高くなるようになっている. すなわち、過去に出現していないものほど順位が高く、逆に過去に何度も出現したものは順位が低くなる.

6.2 観察

以下で述べる事例は、主に算出方法 1 で順位付けを行ったデータを観察した結果である.

6.2.1 流行の全体像の把握

データを観察すると、上位には「フォローありがとう⁶」というような挨拶に関わる単語がほとんどを占める. これはどのような時期にも見られ、ここから **twitter** というツールが計測された 1 年に渡り人々にコミュニケーションツールとして支持されているということが読み取れる.

しかし、そのようないわゆる定型句ではない単語を含むトピックが、それらよりも多くのツイート数を持つ場合もいくつか見られた. このようなトピックは単発的なイベントの発生に関連するものが多く、その時刻における一時的な流行を表しているといえる. これらが出現するとき、矩形同士をつなぐリボンが途切れたり、あるいはその時刻を挟んで全体的に下降・上昇していたりした (図 6.2).

図 6.3(a) は 2011 年 12 月のデータを表現したものの一部である. この図の中央を見ると、上記のような現象が発生していることが見受けられる. これについて詳しく観察してみると、この日 (2011 年 12 月 9 日) はアニメーション映画『天空の城ラピュタ』が地上波で放映され、その実況をするようなツイートが多かったということが分かった. 図 6.3(b) は該当する日のある矩形にマウスカーソルを乗せたときのもの、図 6.3(c) はその矩形をクリックしたときの

⁶あるユーザがそのユーザとつながりを持とうと働きかけたことに対する感謝の意

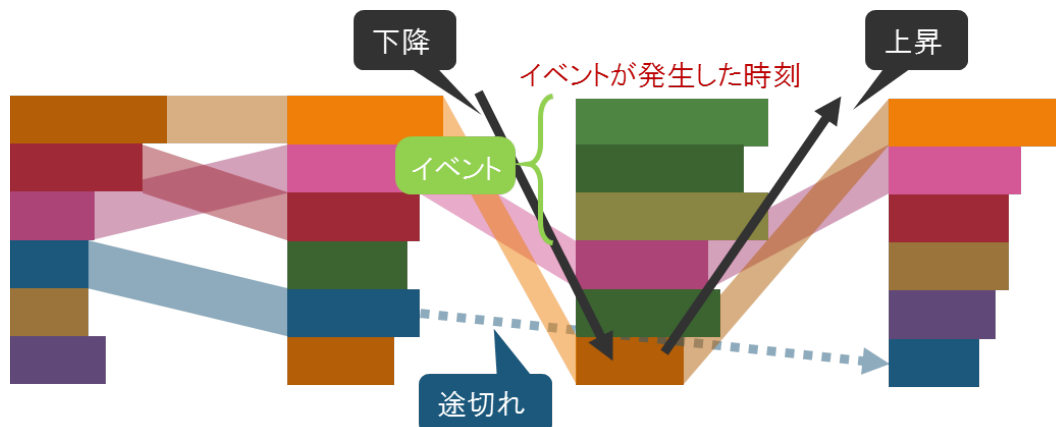


図 6.2: 単発的なイベントの発生

画面下部の表示、および図 6.3(d) は該当する日を選択したときの画面右部のトピック一覧表示である。

同様に、2011 年 7 月 18 日のデータでもこの現象が起こっており、詳しく観察すると、サッカー女子日本代表（なでしこジャパン）が女子 W 杯で優勝したことにに関する単語が上位に登っていた（図 6.4）。

6.2.2 量的値の変動の把握

クリスマスや正月といった季節のイベント、金環日食などの特別なイベントが発生している日は、通常の何も無い日に比べて矩形の横幅が長いものが増えている（図 6.5）。また、twitter 上で行われた「特定のアカウントへメッセージを投稿するとプレゼントがもらえる」というイベントや、ハッシュタグを利用してユーザ同士が交流する行動が盛り上がった時なども、同様にツイート量の増加が確認できた。

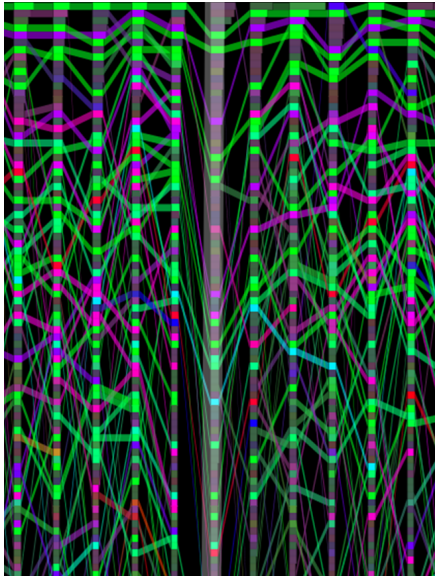
6.2.3 トピックのカテゴリの把握

RT 率が高いトピックをハイライト表示させたときに、比較的緑色が目立つことに気づいた（図 6.6(a)）。なおこの図は算出方法 2 によって順位付けされたデータを適用したものである。

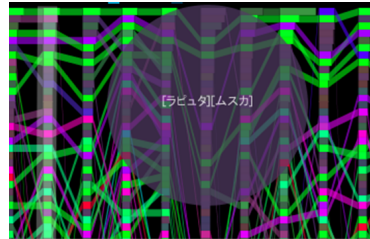
また、個々の色相を観察したときも、12 月 24 日から 25 日にかけては特定の色の矩形が目立った。これらのトピックは「クリスマス」という単語を含むものが多かった（図 6.6(b)）。

6.2.4 データの全体像の把握

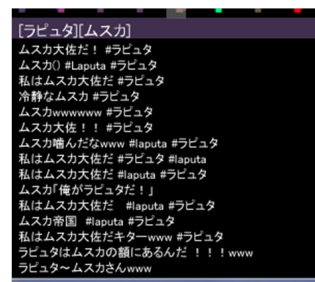
実際に 2011 年・2012 年にどのような出来事があったかという事実と見比べながら分析を行った。その結果、3 月 11 日の東日本大地震や 8 月 15 日の終戦記念日に関するツイートは



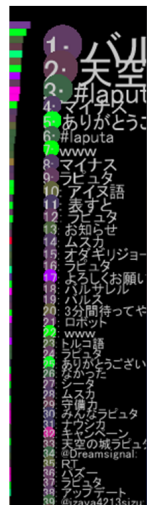
(a) 2011年12月9日のデータにおいて見受けられる谷間



(b) 矩形にマウスカースルを乗せた時の表示



(c) 矩形をクリックした時の表示



(d) 2011年12月9日を選択した時の表示

図 6.3: 『天空の城ラピュタ』に関するトピックの発見と観察

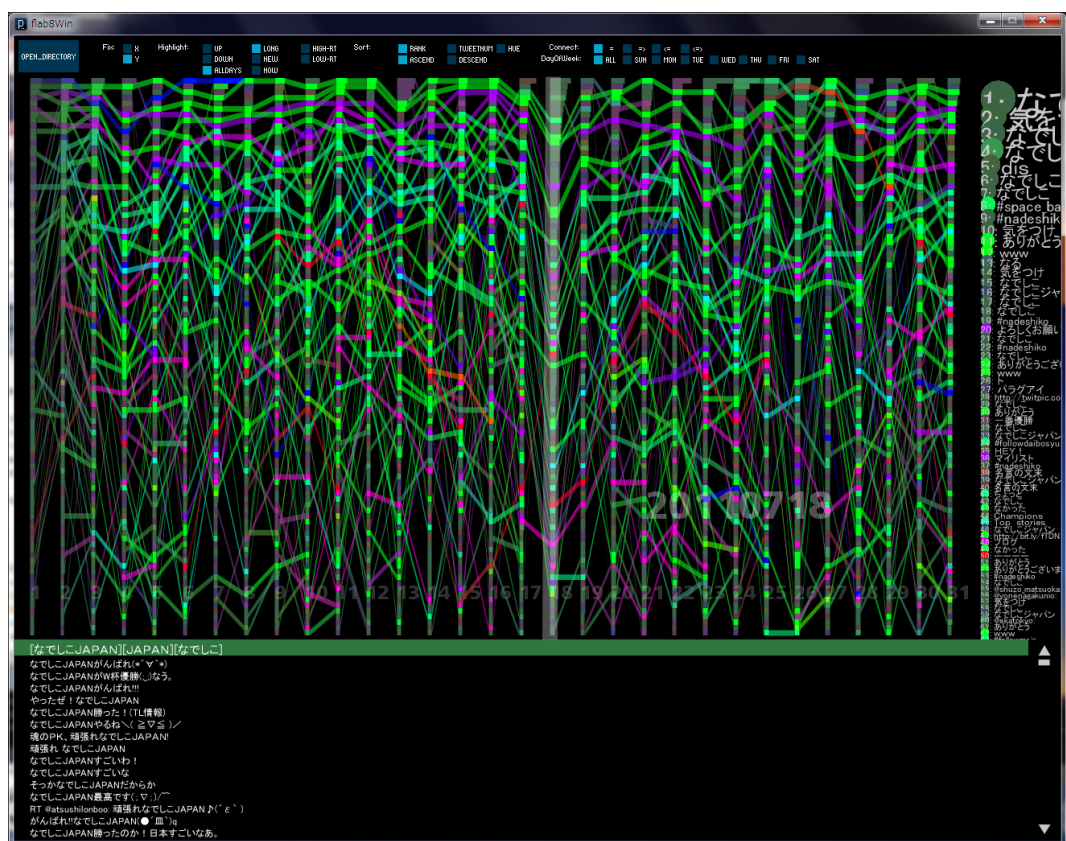


図 6.4: サッカー女子日本代表の W 杯優勝に関するトピックの発見と観察

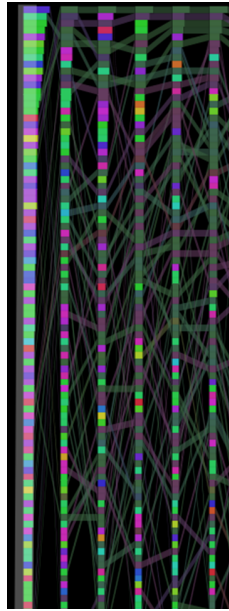


図 6.5: 1 月 1 日（左端）の矩形とその他の軸の矩形の比較

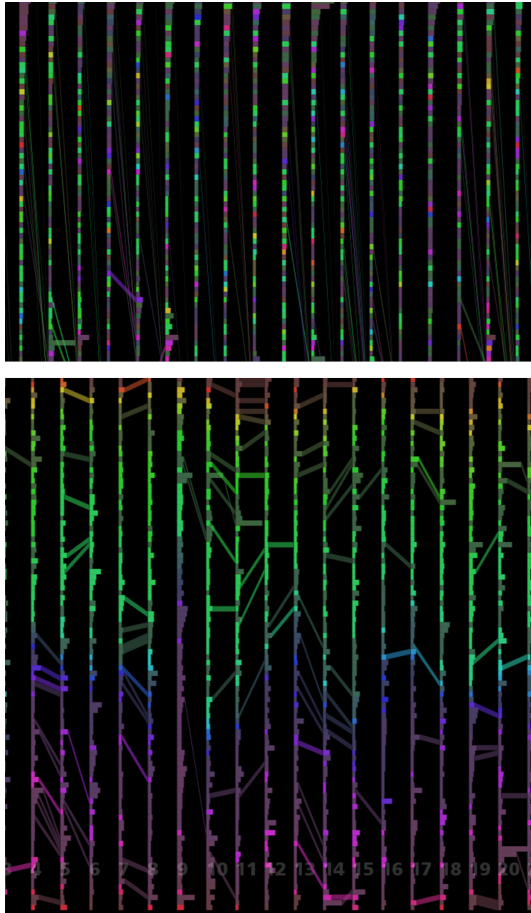
予想よりも少なく，また 2011 年 8 月の民主党代表選，同年 12 月の北朝鮮の金正日総書記死去といった政治的なニュースに関するツイートはさらに少ないと感じた．

反して，2011 年 10 月に Apple のスティーブ・ジョブズ元 CEO が死去したという IT 系のものや，2011 年 11 月 11 日の江崎グリコが提唱する「ポッキー&プリッツの日」，2012 年 5 月の金環日食といったイベント的なものに関するツイートは非常に多かった．

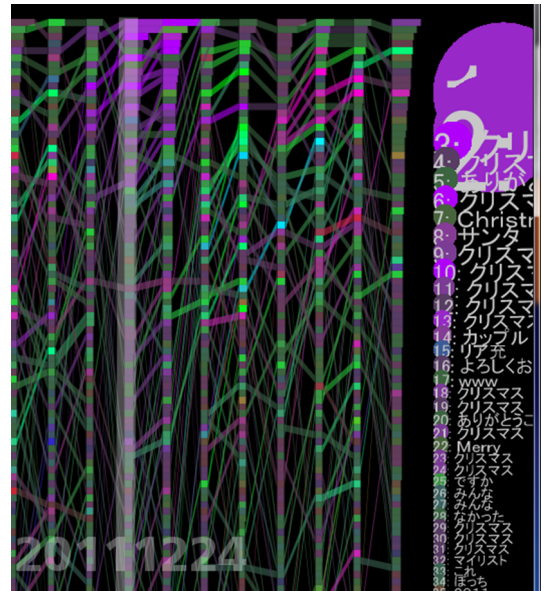
6.3 考察

図 6.3 で観察できた「谷間」を引き起こした現象について調査を行った．その結果，物語中に登場する「バルス」というセリフが，1 秒間における世界中のツイート投稿数 25,088 という当時の新記録を達成するなど，twitter 上でもこの作品のテレビ放映は大きな話題となっていた [26]．

また，6.2.4 節から，twitter 上で流行していた話題は，インターネット上のソーシャルメディアらしくサブカルチャー的なものが多いということが分かった．すなわち，このツールを用いて twitter のトピックデータを分析することで，twitter というメディアにおけるユーザの反応の傾向を得ることができた．



(a) RT 率が高いもののハイライト表示（算出方法 2
によって順位付けされたデータ）



(b) 「クリスマス」を含むトピック群（マゼンタの
もの）

図 6.6: カテゴリの把握

6.4 ツールの有用性に関する議論

観察より，本研究で開発したツールを用いることで，twitter というメディアでの流行について，それぞれの特徴および全体的な特徴を把握することができた．

6.4.1 事象の連続性について

図 6.2 のような現象は，ある時刻において，それまで流行していた事象の順位が下がったり，あるいは事象自体が出現しなかったりしたことを表す．このような表現の「谷間」を探すことは，瞬間的な流行の発見につながる．

個々のトピックがあまりにも長く続いているものは「慢性的な」ものでもあり，挨拶などの頻出するトピックは常に出てくるからと言って流行しているとは言えない．よって，「話題が長く続いているものをハイライト表示する」という機能は，流行の事象を検出するというよりも，そのデータにおけるスタンダードなものを検出するという意味で，データの全体像を把握できる機能だと言える．

逆に，単発的なものをハイライト表示すると，その時刻で瞬間的に流行したのが見える．つまり，その事象が示されたその瞬間にしか興味を持たれないような，継続した話題性がないようなものである．そのような事象がどういった目的で人々に提示されたのかを併せて考えることで，人々の興味関心はどこにあるのかということを探ることができる．

6.4.2 データの傾向把握について

6.2.4 節から，開発したツールでは，トレンドを分析する一方で，twitter というメディアの全体的な傾向というものも把握できた．

メディアの傾向とは，例えば，寄せられたコメントなどを利用し，同じようなデータを新聞社のサイトやニュースサイトからも抽出したとする．すると，おそらく今回のユースケースで得られたような視覚的表現とは違った図が得られ，twitter とこれらのメディアのユーザは性質が違うということが分かるだろう．

すなわち，ツールから提示された結果を分析することで，そのメディアの発信する情報のうちどのようなものがユーザの興味を引いているか，どのようなユーザがそのメディアを好んでいるかといった特性を知ることができる．

6.4.3 テキストラベルからの色相自動算出について

色相の自動算出によるカテゴリ分けについては，6.2.3 節や図 6.6 からある程度はうまくいったと考えられる．

このユースケースで用いた twitter のトピックデータにおいて，カテゴリとは「そのトピックが共通して話題にしている事柄」であると言える．例として図 6.6 で挙げた「クリスマス」について考える．「クリスマスプレゼント」や「メリークリスマス」は「クリスマス」という

単語をトピックのキーワードとしては持っていないが、すべて「クリスマス」という共通の事柄（カテゴリ）に関するトピックである。これらは今回提案した色相の割り当てによって似たような色相で表されており、この手法は有益であると言える。

このように、提案手法においては、平均値から算出することで、多少はトピック内の単語にばらつきがあったとしても、それを補うことができた。しかし、現状は文字列の内容を全く考慮していない算出方法であるため、今回のようなテキストデータに対して用いるにはさらなる改良が必要だと考える。上記の例でいえば、「サンタクロース」や「Christmas」は明らかに「クリスマス」というカテゴリに属するトピックのキーワードであるが、表現としての文字列が「クリスマス」とは全く異なるため、同じような色相を割り当てられない。このような場合、辞書データの参照によるカテゴリ分けなどが改善案として考えられる。

高 RT 率を持つトピックの色相 RT 率が高いトピックが緑色で表示されていたことについて（図 6.6(a)），今回の算出方法では概して緑色で表示される矩形が全体の半分から 1/3 ほどであるのも要因のひとつと考えることもできる。しかし、さらに言えば、RT であるようなツイートを含むトピックの単語には「RT」という文字や、元の発言をしたユーザの ID、あるいは添付されている画像や共有したいと思った Web ページの URL などが含まれることが多い。このような単語はほとんど 1 バイト文字で構成される文字列であるため、色相を算出する際に利用されたのがこれらの単語である可能性が高く、また 1 バイト文字の平均的な値が緑に近い色になる色相値に割り当てられたためだと推測できる。

日本語以外のテキストに対する色相自動算出 今回用いたデータは日本語で記述されたデータ、つまり 2 バイト文字を多く使用するデータである。欧文のような文字数が少ないデータを用いたとき、今回提唱した色相値の算出方法は適しているとはいいがたいと考える。

例として「年ごとの米国の新生児命名ランキング」を可視化するとしたとき、テキストラベルに名前そのものを用いることになる。名前に使用されるアルファベットはたかだか 26 文字であり、この場合名前の 1 文字目が大文字、残りが小文字という基本的な規則があるため、それらを区別する必要はない。つまり、使用される文字数が少なく、日本語のようにカタカナやひらがな、漢字、アルファベットなど字の種類が単語の特徴になるものでもない。このとき値を平均化してしまうと、それぞれが一樣な値になりやすくなり、結果事象の色相間にあまり差が見られなくなってしまうと予測される。

この場合は、例えば名前の最初のアルファベット 3 文字を、大文字・小文字関係なく A から Z に 1 から 26 の数値を割り当て、1 文字目が 3 桁目、2 文字目が 2 桁目…というように、26 進数の 3 桁の数字を利用して色相を算出する手法が考えられるだろう。

第7章 おわりに

7.1 まとめ

時系列量的順位データから、順位と量の両方を時間軸に沿って可視化し、それを利用してトレンド分析を支援するツールを作成した。これは、事象のもつ量よりも順位に着目し、その推移を俯瞰できるような視覚的表現を提供する。かつ、インタラクティブ操作を用いることでより詳細な情報を得られるものである。本ツールを用い、目的に応じて表現を変更することで、トレンド分析に対する要求に対応した。さらに、事象の区別に色相を利用することを考え、その色相値を自動で算出する手法を提示し、この手法の有用性を述べた。

ユースケースとしては、twitter のトピックデータを利用したものを掲げた。このユースケースから得られた結果は、そのデータについてのトレンドがどのようなものであるだけでなく、データ全体の持つ流行の方向性なども観察することができたというものであった。加えて、本研究で作成したツールにより、他のマスメディア、例えば新聞や各ニュースサイトなどからも同様のデータを抽出し、同じように分析することで、各メディアとそれを利用するユーザの特性も把握できるという可能性を示した。

7.2 今後の課題

トレンド分析の要求としては、過去・現在・未来においてのそれぞれの流行を知りたいというものが存在する。今回作成したツールでは、主に過去のトレンドを追うことに適しており、ここから現在および未来のトレンドの分析、あるいは予測ができるようにするためには、まず「流行しそうなもの」がいったいどのようなものであるかということを改めて定義する必要がある。その定義に基づき、何らかのパラメータを可視化することでこの要求に対する支援が行える。

事象のカテゴリ分けも重大な課題である。今回はテキストラベルをもとに色相を算出し、それが近いものをカテゴリ的に近いとみなすようにしているが、この手法では字面が同じものが同様のカテゴリだとみなされ、意味的なことは全く反映されていない。そこで、事象のカテゴリ分けをルールにしたがって正確に行い、それを元に色分けを行うことで、流行の全体像がいつそう明確に分かるようになるであろう。加えて、日本語以外の言語を含むデータを用いた分析をする場合においては改善をすべきである。

表現上の課題としては、順位が極端に離れているもの同士をつなぐリボンが非常に細くなり、どことどこが対応しているかが見づらいというものがある。これについて、曲線を用い

るといった線の引き方を変更する工夫も考えられる。しかし、リボンの傾き具合がランクの上下の幅をより明確にしている場合もあるため、この特性を失わないような改良を考える必要がある。

謝辞

本研究を進めるにあたり、三末和男准教授をはじめ田中二郎教授、高橋伸准教授、志築文太郎准教授にはたくさんの助言をいただきました。その中でも、指導教官として三末先生につきましては特にお世話になりました。時には優しく、またときには厳しくご指導してくださいました。ありがとうございました。

IPLAB の学生の皆様、とりわけ NAIS チームの方々にも心からお礼申し上げます。研究の事のみならず、さまざまな相談にのっていただき、右も左も分からない私を優しくサポートしてくださり本当に感謝しています。

最後に、こうして大学という場で、研究室の仲間と机を並べて学問に励むことができるのも、ひとえに支え続けてくれた家族のおかげです。

感謝してもしきれないほどではありますが、この謝辞をもって感謝のことばと致したいと思います。

なお、本研究の一部は、株式会社富士通研究所からの受託研究によるもので、ユースケースに用いた **twitter** のトピックデータについても提供していただきました。研究の機会を与えてくださったことや、ご多忙中にもかかわらず、本研究についてたくさんの有益なアドバイス、及び質問へのご回答をいただいたことなど、誠にありがとうございました。

参考文献

- [1] Yahoo!検索ランキング. <http://searchranking.yahoo.co.jp/>.
- [2] Sony music online japan : ランキング. <http://www.sonymusic.co.jp/ranking/>.
- [3] Wolfgang Aigner. *Visualization of Time-Oriented Data*. Springer, 2011.
- [4] William Playfair. Commercial and political atlas, 1786.
- [5] Martin Wattenberg. Baby names, visualization, and social data analysis. In *IEEE Symposium on Information Visualization (InfoVis 05)*, pp. 1–7, 2005.
- [6] Joonghoon Lee. Exploring global terrorism data: a web-based visualization of temporal data. *Crossroads*, Vol. 15, No. 2, pp. 7–14, December 2008.
- [7] Susan Havre, Elizabeth Hetzler, Paul Whitney, and Lucy Nowell. Themeriver: Visualizing thematic changes in large document collections. *Visualization and Computer Graphics*, pp. 9–20, 2002.
- [8] Jacques Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, 1983.
- [9] The economist: Cutting red tape. <http://www.economist.com/>.
- [10] William C. Brinton. *Graphic methods for presenting facts*. The Engineering magazine company, 1914.
- [11] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphic Press, 1983.
- [12] Edward R. Tufte. *Envisioning Information*. Graphics Press, 1990.
- [13] The telegraph: The top 100 baby names in england and wales in 2011. <http://www.telegraph.co.uk/>.
- [14] Conglei Shi, Weiwei Cui, Shixia Liu, Panpan Xu, Wei Chen, and Huamin Qu. Rankexplorer: Visualization of ranking changes in large time series data. *Visualization and Computer Graphics*, pp. 2669–2678, 2012.

- [15] Shixia Liu, Michelle X. Zhou, Shimei Pan, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. Interactive, topic-based visual text summarization and analysis. In *CIKM '09 Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 543–552, 2009.
- [16] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. Tiara: A visual exploratory text analytic system. In *KDD '10 Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 153–162, 2010.
- [17] Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. Textflow: Towards better understanding of evolving topics in text. *Visualization and Computer Graphics*, pp. 2412–2421, 2011.
- [18] Edward R. Tufte. *Beautiful Evidence*. Graphics Press, 2006.
- [19] Bongshin Lee, Nathalie Henry Riche, Amy K. Karlson, and Sheelash Carpendale. Spark-clouds: Visualizing trends in tag clouds. *Visualization and Computer Graphics*, pp. 1182–1189, 2010.
- [20] Christopher Collins, Fernanda B. Viégas, and Martin Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *Visual Analytics Science and Technology*, pp. 91–98, 2009.
- [21] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *VIS '90 Proceedings of the 1st conference on Visualization '90*, pp. 361–378, 1990.
- [22] Robert Kosara, Fabian Bendix, and Helwig Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *Visualization and Computer Graphics*, pp. 558–568, 2006.
- [23] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. In *CHI '04 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 575–582, 2004.
- [24] Martin Wattenberg, Fernanda B. Viégas, and Katherine Hollenbach. Visualizing activity on wikipedia with chromograms. *Lecture Notes in Computer Science*, Vol. 4663, pp. 272–287, 2007.
- [25] Ben Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pp. 336–343, 1996.
- [26] Cnet japan: 「天空の城ラピュタ」が世界新記録–1秒間のツイート数. <http://japan.cnet.com/>.