

大量の時刻情報付きデータを分析するための可視化ツール

白井 智子 三末 和男 田中 二郎*

概要. 本研究で扱う時刻情報付きデータとは、行動や事件などのイベントと、それらのイベントが発生した時刻を記録したデータである。このデータを分析すると、行動パターンがわかり、マーケティング分野やセキュリティなどの様々な分野へ応用することができる。本研究では、アナログ時計の文字盤のような円を基本とし、イベントの時刻の集合をその中に配置することで、データに記録されているイベントと時刻の関係性のおおよその分布を視覚的に提示する、ChronoView を提案する。ChronoView では、時刻の集合を位置で表すが、この場合、時刻の集合が同じ位置に配置されてしまう可能性があるという点で、表現上の曖昧性がある。本研究では、この表現上の曖昧性を改善し、イベント発生の特徴がより把握しやすい表現を検討した。

1 はじめに

本研究で対象とする「時刻情報付きデータ」とは、行動や事件などのイベントとそれらが発生した時刻を記録したデータである。時刻情報付きデータには、Web アクセスログ、ショッピングサイトでの商品の購買履歴、Twitter のツイートなど、様々な種類のデータがある。時刻情報付きデータを分析していくと、イベント発生の特徴がわかる。これらの特徴の把握は、マーケティングやセキュリティ分野を初めとする様々な分野で期待されている。

本研究で扱うイベントは、1 つ以上の時刻で発生する。たとえば、ある Web ページが 10 時、11 時に閲覧されたとすると、その Web ページへのアクセスは 2 つの時刻で発生したことになる。時刻情報付きデータを分析するための既存手法 [1] の多くは、1 つの時刻の集合を表示するために一定の領域を必要とする。そのため、大量の時刻情報付きデータを一度に表示することは困難であり、データの持つ特徴を俯瞰により把握することは容易ではない。

本研究では、大量の時刻情報付きデータからイベント発生の特徴分析を支援するため、時刻の集合を二次元平面上の位置で表現する ChronoView を提案する。ChronoView は、アナログ時計の文字盤のような円を基本とした手法である [2]。

2 視覚的表現

2.1 位置を使った時刻の集合の視覚的表現

ChronoView は、アナログ時計の文字盤のような二次元平面上にイベントを配置する表現である。

まず、1 つの時刻を二次元平面上に配置すること

Copyright is held by the author(s).

* Satoko Shiroy, 筑波大学大学院 システム情報工学研究科 コンピュータサイエンス専攻, Kazuo Misue and Jiro Tanaka, 筑波大学 システム情報系

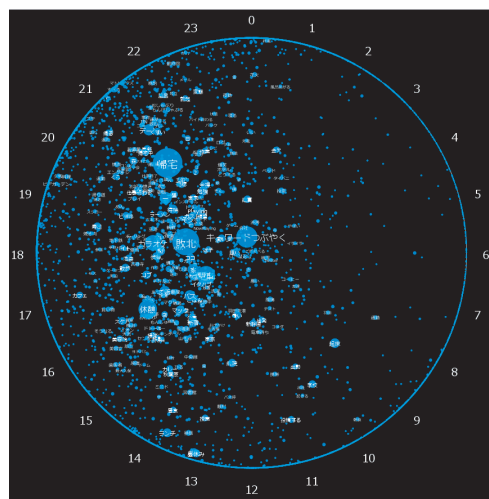


図 1. ChronoView の概観

を考える。時刻が t_0 のとき、アナログ時計の文字盤の 12 時の位置に配置し、半径 r の円周上に時計回りに周期 c で配置する。すべての時刻の集合を U とすると、このような配置は関数 $f_0 : U \rightarrow R^2$ によって次のように定義できる。

$$f_0(t) = (r \cos \theta, r \sin \theta) \quad (1)$$

ただし、

$$\theta = \frac{\pi}{2} - 2\pi \frac{t - t_0}{c}.$$

空ではない時刻の有限集合 $T = \{t_1, t_2, \dots, t_n\} \in 2^U$ は、要素 t_i ($i = 1, \dots, n$) に対する平面上の点の重心に配置する。このような配置は関数 $f : 2^U \rightarrow R^2$ によって次のように定義できる。

$$f(T) = \frac{1}{|T|} \sum_{t \in T} f_0(t) \quad (2)$$

このように、あるイベント e が時刻の有限集合 T を持つとき、 e を $f(T)$ に配置する。

図 1 では、 t_0 を午前 0 時、 c を 24 時間としている。 c を 1 時間とすれば、1 時間周期の図を表現できる。また、 c を 7 日間とすると、1 週間周期の図を表現できる。この配置方法により、イベントが広範囲に分散して発生しているのか、特定の時刻に集中して発生しているのかなど、おおよその特徴把握が可能になる。

2.2 金平糖表現

ChronoView は、異なる時刻の集合が同じ位置を占める可能性がある点で、表現上の曖昧性がある。本研究では、この曖昧性を解消し、複数イベントを一度に比較できる表現として、金平糖表現を考えた。

まず、図 1 において、1 つのイベントに着目し、そのイベントについて、小さなアナログ時計の文字盤のような円を考える。このとき、着目したイベントを、小さな円の中心とする。そして、円の中心から、実際にイベントが発生した時刻の方向へ、直線を描画する。この方法で 1 つ 1 つのイベントを描画することで、様々な形の金平糖のようになる。



図 2. 金平糖表現の一例

図 2 は、金平糖表現を適用した例である。この例では、周期を 24 時間とし、1 時間ごとのイベントの発生頻度を集計し、各時間帯の頻度を線の長さで表現している。そのため、イベントの発生頻度が高い時間帯ほど、線の長さが長くなり、発生頻度が低いほど、線が短くなる。

この表現により、形を見る事で、おおよかな発生時刻の特徴を読み取る事ができる。また、一度に複数のイベント発生の特徴を把握し、比較することが可能である。

3 実行結果

本研究では、「起床なう」、「忘年会なう」のように「なう」というキーワードを含むツイートデータを ChronoView で視覚的に提示した。利用したツイートデータは、ユーザ名、日付、時刻、ツイート文を持つ「なう」を含むツイートを MeCab¹を利用して

形態素解析し、「なう」の直前にある人間の行動にあてはまる単語列を抽出した。今回は、抽出した単語列をイベントとした。利用データは、2011 年 12 月 1 日から 12 月 31 日までの 103,194 ツイートである。

ChronoView より、まず位置からイベントの大まかな特徴が読みとれた。円の淵側を見ると、23 時の付近に「月食」、13 時の付近に「ランチ」というイベントが配置されていた。このことから、これらのイベントは、特定時刻に依存していそうだと、ということがわかった。また、円の中央付近を見ると、「電車」、「バス」、「カラオケ」というイベントが配置されていた。このことから、これらのイベントは、幅広い時間帯に発生していそうだと、ということが読みとれた。

ここでさらに、「電車」と「バス」の金平糖の形状に注目した(図 2)。「電車」は、7 時、8 時、18 時の線が長く、10 時から 17 時までの間にもやや長めの線が引かれていたが、1 時から 4 時の間には線がなかった。一方、バスは、12 時、17 時、18 時、21 時の線が長く、6 時から 11 時と 13 時から 16 時にも線が引かれていたが、0 時から 5 時には線がなかった。このように、イベントの位置が近いもの同士であっても、金平糖表現より、イベントの発生時間の違いを読み取ることができた。

4 まとめ

本研究では、大量の時刻情報付きデータからの特徴分析を支援するための表現手法である ChronoView を提案した。ChronoView では、時刻の集合をアナログ時計を模した二次元平面上の位置で示すことで、イベントと時刻の関係性を提示した。

謝辞

本成果の一部は、株式会社富士通研究所からの受託研究によるものである。

参考文献

- [1] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann and Christian Tominski: *Visualization of Time-Oriented Data*, Springer (2011).
- [2] Satoko Shiroi, Kazuo Misue and Jiro Tanaka, ChronoView: Visualization Technique for Many Temporal Data, *16th International Conference Information Visualization (IV2012, July 11-13, 2012, Montpellier, France)*, p112-117, (2012).

¹ <http://mecab.sourceforge.net/>