

# 属性を付属したテキスト可視化のための属性付き KeyGraph

## A Feature-attached KeyGraph for the text visualization which was attached in the attribute

伊藤貴一\*1  
Takaichi Ito

諏訪正樹\*1  
Masaki Suwa

大澤幸生\*2  
Yukio Ohsawa

\*1 慶應義塾大学  
Keio University

\*2 東京大学大学院工学系研究科  
School of Engineering, The University of Tokyo

We have devised a feature-attached KeyGraph; which nodes and links are frequent and strong, and how it changes over different periods are visualized and easily interpretable by users. We have applied this tool to the meta-cognition of a baseball player. By looking into the change of structure of nodes and links in the feature-attached KeyGraph, we found it possible to interpret how meta-cognition has changed over half a year during which his batting average has drastically changed.

### 1. はじめに

世の中には多くの文章があるが、文章には本来的に、その文章を書いた人、書いた時間などの情報が付属している。単に文章の文字を分析するのではなく、どんな人が、どんな時に書いたものなのか？という情報を含めて、分析することへの社会的需要が増している。例えば、書いた人の違いを含めて分析することで、立場の違いによる意見の違いの分析や、同じ人の書いた文章を時系列で分析することで、意識の変化を分析することができる。このことは、社会分析、市場分析、認知分析のうえで重要なことであり、分析する手法が求められる。この分析において必要なことは気付きを誘発し、直感的なイメージが描きやすいということである。

### 2. 属性付き KeyGraph

気付きを誘発するテキストマイニング手法として KeyGraph[1]がある。本論文では、KeyGraph の機能拡張した、属性付き KeyGraph[伊藤 2008]について論じる。

属性付き KeyGraph とは、元々アンケートなどの自由回答文と、回答者属性をクロスして分析したいという発想から作ったものである。文章に属性をつけたデータを用意し、通常の KeyGraph を作成する。そして、ノードの中に属性ごとの頻度の割合を円グラフで書き込む。ここでいう属性とは、文章に付随するものであり、たとえば、文章を書いた人の性別・年齢・ポジションや、書いた日付などであり、文字列として扱う。日付のような連続値は、対象のすべての日付を2つまたは3つの期間に分けた分類したものを上位の属性とし、可視化する。そうすることで、複数の時間的変化の量的変化を可視化することができる。

#### 2.1 属性リンクの生成

[伊藤 2008]において、課題とした、量的差異の表現ができていないが、構造的な差異が見えないという問題に対して、属性リンクを可視化することで解決をはかった[伊藤 2009]。属性リンクとは、KeyGraph により計算されたノードを元に、各属性毎に集計したデータに基づいて、共起関係を計算し、その関係の上位のものを指定数描画するものである。

一度に多くのリンク関係を可視化すると、情報量が多すぎるの

で、リンク数0から少しずつ操作しながら見せるような仕組みを作成した。副作用として、リンクの強さを把握しながら見るようになってきた。「全期間のキーグラフの上に、各属性において強いリンクだけを表示する」という手法が、各分類ごとの特徴を際立たせ、更に分類間の比較を理解しやすくする。また、属性毎に表示の ON/OFF を実装し、属性数、リンク数が増えすぎたときにも見やすくした。

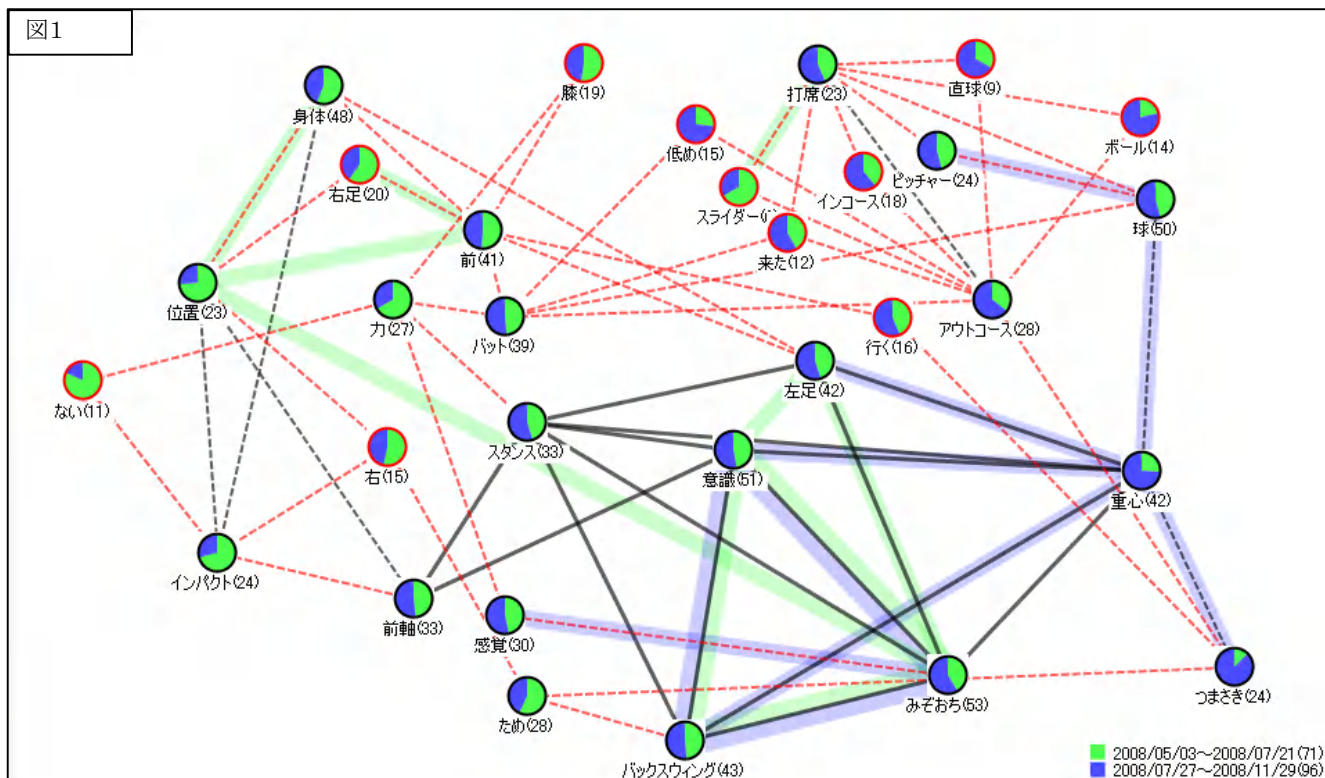
手法をまとめると次のようになる。

- 文章に書いた人の性別・年齢・ポジションや、書いた日付などの属性を付与したデータを用意する。
  - 日付などの連続値の場合は2つ3つに分けて纏めたものを上位の属性とする。
- 属性を無視して通常の KeyGraph を作る。
- 各ノードの中に属性ごとの頻度の割合を円グラフを書き込む
- 表示している KeyGraph に含まれるノードの中から、各属性ごとにまとめた共起関係算出し、それを元に、リンクを指定個数描く。
  - 情報の多さを克服するために
    - ◇ リンク数0から徐々に見せる仕組みを実装した。
    - ◇ 属性毎の表示の ON/OFF を実装した。

### 3. 結果・分析

分析対象として、第2著者が野球のバッティングについて書き溜めたメタ認知の記録を使用する。被験者(第2著者)はそれまで打率 1 割台のスランプだったのを脱出し、2008 年夏に 5 割台にもいほど成績が伸びた。この間に書かれたブログの記事数の合計は 44 日分の日記である。被験者はこの期間、フォームの改造を試みていた。このデータを、日付を前期・後期の二つに分けて、スランプ時、好調時の分析をした。分析の詳細は[伊藤 2009]に詳しい。

黒ノード数 20、黒リンク数 20、赤ノード数 10 リンク指標は Jaccard で作成した、通常の KeyGraph (図の 2) と属性付き KeyGraph (図の 1) である。円グラフの緑が前期、青が後期である。ノードの配置は、割合を元にソートし配置し、バネモデルで計算している。そのため、左から前期の割合が大きいもの順になっている。前期・後期を言うものを付加したため、割合の違いという意味がノード(言葉)に付与された。また、属性リンクを 10 個描画している。これにより、前期と後期では、データの構造が違うことが読み取れる。



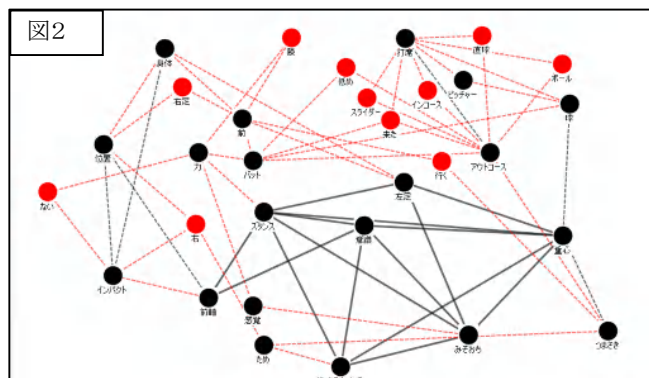
### 3.1 属性リンクの解釈

後期ではリンクは大方、黒リンクになっているものばかりであり、一方、前期のリンクは、黒リンクが発生していないものが多い。そもそも、黒リンクが発生するのは、データ全体で関係性があるものである。指標は  $Jaccard(A \cap B / A \cup B)$  を使っているため、A と B の頻度の差が小さくて(ともに大きくて)共起関係が強いものの評価が高くなる指標である。そのため、当たり前のような関係を指し示すことが多い。つまり、黒リンクは前期・後期合わせた形での被験者の頭の中にある「常識的」な構造である。そのため、黒リンクと齟齬が発生してしまうリンクは、全体として考えていたこととは違うことを考えていたことである。前期に現れるものなら、前期に考えていたが、後に棄却された関係であり、後期に現れるのなら、新たに考え始めている関係である。

そういう観点でみると、「位置-みぞおち」「位置-前」「意識-左足」という関係は前期には頻繁に考えていたがその後被験者の中で捨てられたものである。一方、後期は、すべてが黒リンク、赤リンク(同じ指標を使っているので準黒リンクといえる)の上にある。つまり、前期に強く意識していた「位置」周辺などの「緑リンクのみの黒リンクなし」関係を捨て去り、前期にも頭の片隅にあった関係(もしくは全く関係を見出していなかったものの間の関係) = 「青リンクで黒 or 赤リンク」関係を考えるようにメタ認知意識を再構築し、身体の動かし方に関する被験者の意識が統合されて来たといえるだろう。また、これは、印象として、前期が放射状の開いたリンクが多く(発散)、後期は多角形に閉じたリンクが多い(収束)というように、発散のフェーズから収束のフェーズに移行したとも言い表せるだろう。

### 4. まとめ

以上論じたように、時期ごとのリンクの位置や構造を比較することにより、8月以降(上の分析では後期に合致する)の成績がスランプを脱却して5割以上の打率を保ったという事実が、言葉の分析ではこのような形で現れているのである。



今回示したのは、日付を属性とし、時間変化による一人の人の認知の変化を可視化したものである。図2のオリジナルのKeyGraphでは見えない関係について理解できることがわかるだろう。時間以外にも、性や年齢などの基本統計情報や、立場などの情報を付加して、多人数における立場の違いによる、言葉の量的変化、構造的変化の可視化も同様可能である。

### 参考文献

[大澤 2003]大澤幸生:チャンス発見の情報技術—ポストデータマイニング時代の意志決定支援, 東京電機大学出版局(2003)  
 [伊藤 2008]伊藤貴一,大澤幸生:属性付きキーグラフ,第 22 回人工知能学会全国大会, 2B2-04, (2008)  
 [伊藤 2009]伊藤貴一,諏訪正樹,大澤幸生:メタ認知を促進するツールとしての属性付き KeyGraph 分析, 人工知能学会第二種研究会「身体知研究会」第三回研究会, (2009)