テキストマイニングツールを利用したIT系ニュース記事の分析

担当者：若山 優一　指導教員：長田 茂美　教授

1．はじめに

近年，ビックデータは人類の財産と言われ，ビックデータのブームが到来している．企業活動や公的活動などで蓄積された膨大なデータを素早く分析し，問題を解決する分析者のことをデータサイエンティストと呼ぶ．本研究では，IT 系ニュース記事に対して共起ネットワーク分析を適用し，その有効性を検討する．

2．共起ネットワーク

共起ネットワークとは，テキストの中で用いられた単語をノードとし，単語と単語の共起性をリンクとするネットワークであり，リンクの強さを Jaccard 係数で表している．Jaccard 係数 sim(v，q)とは集合間の類似度であり，式(1)により定義されている．

$$sim(v,\ q)=|\,v\cap q\,|/|\,v\cup q\,| \qquad (1)$$

ただし，v，q はテキストにおける 2 単語の出現頻度である．

3．分析方法

IT系ニュース記事としてITmediaのH25年1月分の記事を分析対象とした．分析の方法としては，テキストマイニングのためのフリーソフトウェアであるKH Coderの共起ネットワーク分析を用いた．

ITmediaのニュース記事を3日毎，5日毎，10日毎にテキストファイルにまとめたものを分析単位とした．KH Coderの共起ネットワーク分析用いるJaccard係数の閾値は0.2に設定した．

4．分析結果

図 1 に，H25 年 1 月 26 日~31 日の 5 日分の分析の結果得られた共起ネットワークと，媒介中心性の役割を担うノードの検出結果を示す．媒介中心性とは，そのノードがアクターとアクターとを媒介している程度を表す指標である．ここで，アクターとは，ノード同士のリンクが強く結びついたノードの集合体である．

KH Coder では，媒介中心性が高い順にピンク，白，水色で表示される．KH Coder では，アクターの検出はできないため、分析者であるデータサイエンティストが，媒介中心性の役割を担うノードに繋がり，かつ，ノード間のリンクの強さが強い（Jaccard 係数の値が高い）ノードの集合体をアクターとして判定する．

図 1 の分析結果からは，「Samsung」，「LINE」，「JAVA」，「カメラ」が媒介中心性の役割を担うノードとして検出されている．

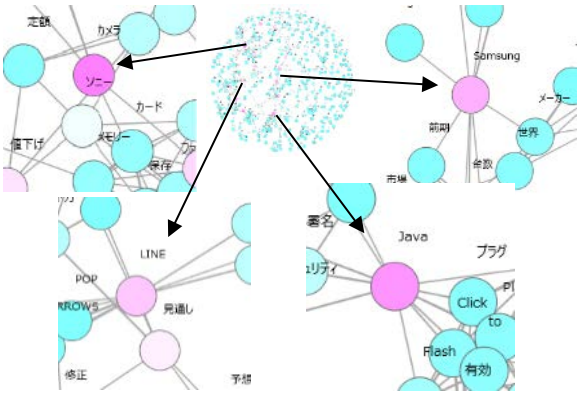

図１．共起ネットワーク分析（媒介中心性）

図 2 に，H25 年 1 月 19 日~21 日の 3 日分の分析の結果得られた共起ネットワークと，サブグラフの検出結果を示す．サブグラフとは，ノード同士のリンクの強さが強い（Jaccard 係数の値が高い）ノードの集合体（アクター）である．KH Coder では，サブネットワークが自動的に色分けして表示される．図 2 の分析結果からは，色分けされたアクターは大きなトピック分けを行っていることが分かる．

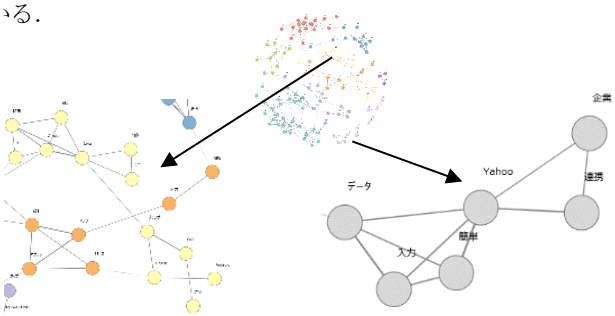

図２．共起ネットワーク分析（サブグラフ検出）

5．まとめ

本研究では，KH Coder の共起ネットワーク分析を IT 系ニュース記事に適用し，その有効性を検討した．媒介中心性の役割を担うノードの検出結果から，共起ネットワークにおける媒介中心性の役割を担う重要なノードは，IT 系ニュース記事に出現する共通の単語であり，IT 市場における注目度を判断できることが分かった．また、サブグラフの検出結果から，検出された複数のサブグラフ（アクター）の色分けが大きなトピック分けとなっていることを発見することができた．その結果，共起ネットワーク分析は IT 系ニュース記事に対して有効であると考える．

6．参考文献

［1］ KH Coder.　http://khc.sourceforge.net/
［2］ 吉見憲二，樋口清秀．共起ネットワーク分析を用いた訳あり市場の考察．　p31-p38　（2012）