5N-1

FACT-Graph を用いた社説の比較分析

高見澤聖子[1]　佐賀亮介[1]　辻洋[2]　松本一教[3]

神奈川工科大学情報学部情報工学科[1]
大阪府立大学大学院工学研究科電気・情報系専攻[2]
神奈川工科大学大学院工学研究科情報工学専攻[3]

1. はじめに

情報洪水時代である近年、情報へのアクセスが容易な一方、膨大な量の情報の取捨選択は利用者の負担を強いる場合もある。しかし、情報を比較し特徴を示すことで、情報の取捨選択などの意思決定支援や情報の有効活用につなげることができ、利用者にとって有益な結果が得られる。

本研究では、トレンド分析に用いられる FACT-Graph(Frequency And Co-occurrence-based Trend Graph)[1]を利用したテキストデータの比較分析法を提案し、その有用性を考察する。

2. FACT-Graph

FACT-Graph は時系列テキストデータを 2 期間に分け、その間の大局的なトレンドを見ることができる(図 1)。これは、情報利用者の意思決定支援や新たな知識発見が得られることを目的としている。

この FACT-Graph には二つの要素技術がある。一つ目はクラス遷移分析[2]である。これはテキストデータ中に存在する単語の重要度が、時系列の変化に伴って重要度も変わる様子を分類したものである。ここで、単語の重要度は TF(Term Frequency)と DF(Document Frequency)によって決められている。そして、その単語を TF と DF の値により、表 1 の Class A から Class D に分類し、ある期間におけるクラスの変化に対して表 2 の意味を持たせることで単語のトレンドを把握する。また、表 1 の Pattern1 は表 2 における「Hot」から「Negligible」の対角線要素を表し、Pattern2 は対角線要素から左下の 5 項目、Pattern3 は対角線要素から右上の 5 項目をそれぞれ表している。

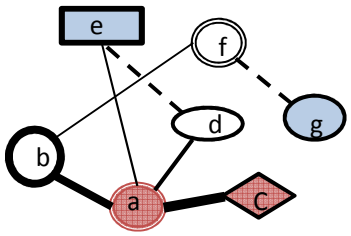

図 1　FACT-Graph のイメージ

表 1　FACT-Graph の要素と属性

	Class A : High TF & DF (Major Words)
	Class B : High DF & Low TF (Complementary Words)
	Class C : Low DF & High TF (Domain Words)
	Class D : Low TF & DF (Minor Words)
	(a) : Continues
	(b) : Occurs
	(c) : Goes Off
	Pattarn1 : Stable
	Pattern2 : On the Rise
	Pattern3 : On the Decline

表 2　クラス遷移分析

$t_1 \rightarrow t_2$ ＼ $t_2 \rightarrow t_3$		DF High		DF Low	
		TF High	TF Low	TF High	TF Low
DF High	TF High	Hot	Cooling	Bipolar	Fade
	TF Low	Common	Universal		Fade
DF Low	TF High	Broaden		Locally Active	Fade
	TF Low	New	Widely New	Locally New	Negligible

二つ目は時系列の変化に伴った共起情報の意味づけ[1]である。共起とは、二つの単語同士がどの程度の頻度で共に出現しているかを表している。その共起関係に対し、時間が経過するにつれ発生または消滅、もしくは継続した関係だと意味づけを与える（表 1 の(a)～(c)）。

FACT-Graph の可視化結果は次の手順より生成される。

1. データを前期と後期の期間に二分割にする。
2. 期間ごとに単語を抽出しクラス分類を行う。
3. 前期と後期の単語を統合し、全単語の共起度とその変化を求める。

Comparison analysis of editorial article by using FACT-Graph
1 Seiko Takamizawa, 1 Ryosuke Saga 2 Hiroshi Tsuji
3 Kazunori Matsumoto
1 Kanagawa Institute of Technology, Faculty of Information Technology
2 Osaka Prefecture University, Graduate School of Engineering
3 Kanagawa Institute of Technology, Graduate School of Engineering

4. クラス遷移分析を用いて、単語の重要度の変化を求める。
5. 単語をノード、共起をリンクとみなし、可視化を行う。

3. FACT-Graphを用いた比較分析法

FACT-Graph における期間という考え方は、視点を変えると 2 カテゴリ間の比較分析を行っていると見なすことができる。この視点から、時系列データではなく比較対象のテキストデータを適用することで、目的とする比較分析が行えると考えられる。また可視化結果を単語の発展や衰退ととらえるのではなく、比較対象の特徴と見なすことで、比較分析として FACT-Graph を利用できるようにする。

比較分析として FACT-Graph を用いるために、単語のクラス遷移分析の解釈を表 3 のように変更した。また比較分析においては、単語の存在の有無は特徴として重要な意味を持つため、新たに Class E を設定した。この Class E はグラフにおいて、丸の破線で表される。さらに、表 1 の(a)～(c)において、共起関係の線を太線から実線、細線と破線をパターンの色と合わせることで、視覚的に特徴を捉えやすくした。

表 3　比較分析のためのクラス遷移分析

Target A \ Target B		DF High		DF Low	
		TF High	TF Low	TF High	TF Low
DF High	TF High	Hot	A : Important B : Universal	A : Important B : Locally Active	A : Important B : Negligible
	TF Low	B : Important A : Universal	Universal		A : Universal B : Negligible
DF Low	TF High	B : Important A : Locally Active		Locally Active	A : Exclusive B : Negligible
	TF Low	B : Important A : Negligible	B : Universal A :Negligible	B : Exclusive A : Negligible	Negligible

4. ケーススタディ

有用性の確認を行うため、FACT-Graph を用いて比較分析が行えるかを検証した。

本ケーススタディでは 2006 年-2008 年の毎日新聞 2123 件、朝日新聞 2033 件の社説の中から、それぞれ「五輪」が含まれている 64 件と 74 件を用いた。表 3 の Target A を毎日新聞、Target B を朝日新聞とすると表 1 より Pattern3 と(c)は毎日新聞、Pattern2 と(b)は朝日新聞に特徴があるものと解釈できる。

対象データを用いた出力結果を図 2 に示す。図 2 において、Pattern1 で Class A のノードは北京五輪と関連したものが多く、話題の中心であることがわかる。また同図の中心部は Pattern3 の Class D または Class E が多く現れており、それらの多くは(c)の関係で繋がっている。さらに、ノードには五輪の競技や選手の名前が書かれている。これらより毎日新聞は五輪の本質的なことについて書いているとわかった。

一方、同図左上部は中国関連、右端は東京都関連のノードが目立つ。これは Pattern2 の Class D で多く現れており (b)で繋がっている場合が多い。これより朝日新聞は「五輪」と政治を絡めて書くことが多いとわかった。

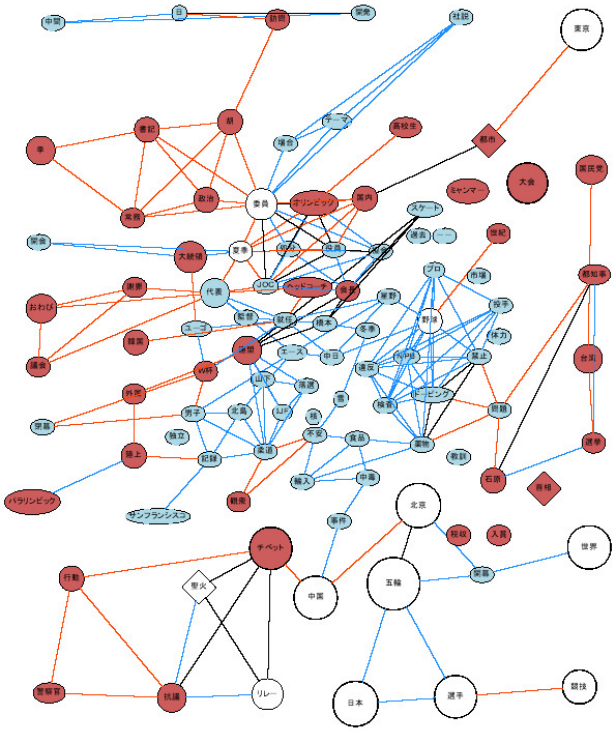

図 2 FACT-Graph を用いた出力結果

5. おわりに

本論文では FACT-Graph を用いたテキストデータの比較分析法を提案した。ケーススタディでは二つの新聞社間の社説をその比較分析法により分析し、二社間の特徴が確認できた。今後の課題として、3 つのカテゴリを一つのグラフに表現すること、比較分析用のクラス遷移分析を精査していくことが挙げられる。

参考文献

[1] 佐賀亮介 他：FACT-Graph：頻度と共起度を用いたトレンド可視化，電気学会論文誌 C，Vol. 129，No. 3，pp. 545-552 (2009.03).

[2] M. Terachi et al.: “Trends Recognition in Journal Papers by Text Mining”, IEEE International Conference on Systems, Man & Cybernetics(IEEE/SMC 2006), pp. 4784-4789(2006).