3B3-NFC-01b-3

FACT-Graph と TETDM の融合の可能性

Speculation with regard to combination between FACT-Graph and TETDM

佐賀　亮介[*1]
Ryosuke Saga

[*1] 大阪府立大学工学研究科
Graduate School of Engineering, Osaka Prefecture University

This paper describes the speculation with regard to combination between FACT-Graph and TETDM. TETDM has been developed for environment of text mining by trying and comparing several text mining methods. On the other hand, FACT-Graph also has been developed for visualization method. This paper consider the possibility of embedding FACT-Graph in TETDM.

1. はじめに

テキストマイニング技術は成熟してきており，様々な機能がユーザに提供され始めてきた．主なテキストマイニングのタスクは，キーワード抽出，トピック抽出，要約，分類，可視化など多岐に渡っている．データサイエンティストという言葉が生まれている昨今において，これらのタスクを解決するためのツールの必要性は増すことが予想される．

このような背景のもと，TETDM (Total Environment for Text Data Mining：テトディーエム)は，人工知能学会における近未来チャレンジテーマの一つとして採択され，TETDM 統合環境として開発がなされている[砂山 11]．様々なテキストマイニング手法を組み合わせた統合開発環境を目指しており，多くのモジュールからなることが多い．

一方，佐賀らは FACT-Graph と呼ばれる可視化手法を開発している．FACT-Graph は複数の異なる次元の情報を一つにまとめることを目的に開発されており，様々な応用事例がある．

本論文では，FACT-Graph と TETDM について考察し，TETDM に FACT-Graph をモジュールの一つとして組み合わせることによって，どのような取り組みができるかどうかについて，考察する．

2. TETDM と統合環境

TETDM はテキストマイニングにおけるキーワード抽出や共起，可視化など様々なタスクを実行し，比較検討や機能拡張が可能になることを目指したプロジェクトである．それを実現するための統合環境の開発が，現在有志により進められており，タグクラウド生成・可視化などの実装が進んできている．これらの環境はそのプロジェクト名から TETDM 統合環境，または TETDM と呼ばれている(以降，TETDM と呼ぶ)．

この TETDM は，マイニング処理モジュールと可視化インターフェイスモジュールから構成されている．後者のモジュールは前者のものを出力するためのものであり，共通的なインターフェイスにより制御されている．

これらの詳しい内容・詳細は[砂山 11][高間 11]を参考にされたい．

3. FACT-Graph

FACT-Graph とは共起グラフを基にした情報可視化法とその可視化結果自体を指すものである[佐賀 09]．FACT-Graph のコンセプトは，複数の異なる次元の情報を一つにまとめることを重要としており，いままで，複数の可視化結果を用意し，比較していたものを一つにまとめることを目的としている．

最初の FACT-Graph はテキストデータのトレンド分析手法として開発された．この FACT-Graph を用いることで，キーワードの状態と共起の状態だけでなく，それらのトレンドがどのように変化しているかを表現可能になった(図 1)．この FACT-Graph により新聞記事や犯罪記事などのテキスト分析が行われている．また，その FACT-Graph の特徴を生かして，Web アクセスログのトレンド可視化[Saga 11_1]やゴルフサイトの可視化分析，また，社説の比較分析[Saga 11_2]，Call For Paper の可視化[Issertial 12] がなされている．

FACT-Graph の本質的考え方と独自性は，様々なデータセットを表す共起グラフ（Temporary Graph）の比較にある． ある Temporary Graph における頂点と枝が何らかのクラスに属していると定義する．そして，複数のデータセット間における頂点や枝の変化量に基づいて状態の変化を表す．この Temporary Graph が時系列に基づいて生成されるとき，その結果として出力される FACT-Graph はトレンドを可視化したものとなり，複数のカテゴリ間に基づいた場合，カテゴリ間の比較分析結果として出力される．そして，この時系列情報や比較分析での特徴は，FACT-Graph の頂点や枝に色として表される（基本的には赤・青・白）．

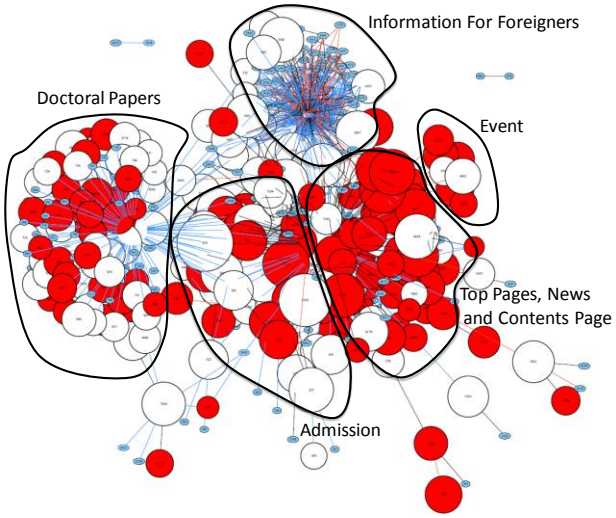

図 1　FACT-Graph による Web アクセスログの可視化

連絡先：連絡先：佐賀　亮介，大阪府立大学工学研究科，大阪府堺市学園町 1−1，saga@cs.osakafu-u.ac.jp

昨今では，より抽象的に可視化できるように，オントロジやカテゴリを組み込むような仕組みや，KWIC 環境やインタラクティブな分析環境を提供したソフトウェアも開発されている[Saga 10]. また，一般的に無向グラフで表される FACT-Graph を有向グラフへと発展し，アソシエーションルールなど可視化できるようにした FACT-Graph も開発されている[Saga 12_4].

4. TETDM への FACT-Graph の組み込み

FACT-Graph の出力プロセスと TETDM の環境は類似点が多い. 図 2 は 2 期間のトレンドを可視化するための FACT-Graph の出力プロセスである. このキーワードとリンクの分析部分，最終的に出力する部分は，それぞれ TETDM のマイニング処理モジュール，可視化インターフェイスモジュールの一部とみなすこともできる. この点から少しずつ融和が可能であると考えている.

一方で，FACT-Graph と TETDM の大きな相違点は，複数の結果を一つにまとめることを目的とするか，ソフトウェアの機能として複数の分析機能を提供することを目的としているかという点である. このため，FACT-Graph は複数の共起グラフの結果の統合など，要約や統合として利用できる可能性がある. 一方で，FACT-Graph 自身には詳細なテキストマイニングプロセスを提供していない. その点で，TETDM において FACT-Graph を用いることは，TETDM の詳細なテキスト分析機能を利用できるというメリットがあると考えられる.

ただし，FACT-Graph を TETDM にて使うためにはいくらかの問題がある. 一つは，データの形式である. FACT-Graph はテキスト情報だけでなく，時間情報が必須となる. TETDM はこの時間情報をまだサポートしていない. また他の問題点はデータの格納法である. FACT-Graph はデータベース上での処理を基本としており，現在，SQLite 上で動作している. これは，FACT-Graph が大量データを取り扱うことを前提としており，オンメモリ上での処理を意図して避け，大量のデータを処理・分散することを構想しているからである. 実際に，Web アクセスログやゴルフサイトの可視化分析では，一千万件以上のデータの可視化に取り組んでいる. しかしながら，TETDM は Java をベースとした環境であり，基本的にオンメモリ上での開発となっている. そのため，ビッグデータのような大量データを扱うことには不向きであるため，分析の際のデータサイズに制限が生じる. その他の問題点は，TETDM ではグラフ構造をプリミティブに扱うことができない点である. FACT-Graph では Temporary Graph の考え方を用いており，結果もすべて Temporary Graph として出力される. すなわち，FACT-Graph は FACT-Graph 同士を統合することを構想しているが，TETDM ではプリミティブに対応していない. これらをうまく実装することが FACT-Graph を組み込むうえで重要となると考えられる.

5. おわりに

本論文では，FACT-Graph と TETDM のコンセプトを考察し，方法論や分析プロセスなどから FACT-Graph を TETDM へ融合させるための可能性を考察した. 可視化手法である FACT-Graph は，可視化結果からユーザに様々な仮説を立てさせることが本質的な目的であり，TETDM とコンセプトなどが共通している部分が多々ある. 一方，仕様の問題などからそのまま組み込むことが難しいといったことも考えられる.

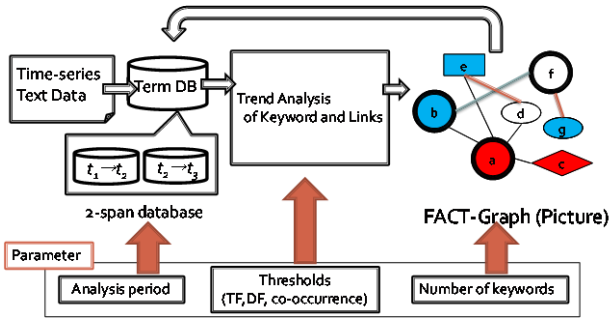

図 1　FACT-Graph 出力プロセス

参考文献

[砂山 11] 砂山 渡, 高間 康史, ダヌシカボレガラ, 西原 陽子, 徳永 秀和, 串間 宗夫, 松下 光範 : Total Environment for Text Data Mining, 人工知能学会論文誌, Vol.26, No.4, pp.483-493, 2011.

[佐賀 09] 佐賀亮介, 寺地雅弘, 辻洋: 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, 129, 545–552, 2009.

[高間 11] 高間 康史, TETDM を利用したタグクラウド生成・可視化ツールの開発,第 6 回 TETDM&情報編纂研究会, pp. 29-32, 2011.

[Saga 10] R. Saga, H. Tsuji, T. Miyamoto, K. Tabata : Development and case study of trend analysis software based on FACT-Graph, Artif. Life Robot., Vol. 15, No. 2, pp. 234-238, 2010.

[Saga 11_1] R. Saga, T. Miyamoto, H. Tsuji, K. Matsumoto: FACT-Graph in Web Log Data. Knowledge-Based and Intelligent Information and Engineering Systems. pp. 271-279. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[Saga 11_2] R. Saga, S. Takamizawa, K. Kitami, H. Tsuji, K. Matsumoto: Comparison Analysis for Text Data by Using FACT-Graph, LNCS 6772, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 75–83, 2011.

[Saga 12_3] R. Saga and H. Tsuji: Comparison Analysis for Text Data by Integrating Two FACT-Graphs, Intelligent Interactive Multimedia: Systems and Services Smart Innovation, Systems and Technologies, Vol. 14, pp. 143-151, 2012.

[Saga 12_4] R. Saga: Proposal of Directed FACT-Graph Based on Association Rule, Proc. of Asian Conference on Information Systems (ACIS), pp. 123-126, 2012.

[Issertial 12] L. Issertial, H. Tsuji, and R. Saga: Visualized Comparison for CFP Datasets by Structure Identification and Ontology, 2012 International Conference on New Trends in Information Science, Service Science and Data Mining, pp. 365-370, 2012.

[佐賀 13] 佐賀亮介, マウリシオ レテリエル, 開作直樹, 高山幸大, 辻 洋: FACT-Graph と逐次確率比検定を用いた Web アクセスログの分析, オペレーションズリサーチ, Vol. 58, No. 2, pp.87-92, 2013.