RF-001

FACT-Graphを用いたトレンド分析支援ツールの開発

Development of Knowledge Discovery Software based on Fact-Graph

佐賀　亮介†　　辻　洋‡　　田畑　邦晃†
Ryosuke Saga　Hiroshi Tsuji　Kuniaki Tabata

1. はじめに

昨今、ストレージの大容量化とネットワークの高速化に従い、組織では大量のテキストデータを保持することが可能になってきている。テキストマイニングは、この背景の中、蓄えられたテキストデータの有用な利用手段として注目を浴びており、非構造な大量のデータであるテキストデータから有用な知識抽出に用いられる。具体的にキーワード抽出[1][2]や要約[3]、アンケート解析[4]、可視化[5]といった分野で利用が進められている。

本論文では様々なテキストマイニングのうち、時系列データを利用したトレンド分析に焦点を当てる[6]。トレンド分析においては、テキストデータ内の語句やトピック（いわゆる、bag-of-words）が用いられる。そのとき、分析者にある分析期間においてどのような語句が注目を浴びているかの理解を支援することが必要である。そのためのツールとして著者らはFACT-Graphを開発した[7]。FACT-Graphは語句の重要性を元にしたクラス遷移分析と、語句の関連性の２つの要素を元とした手法である。この２つの要素が時が移るにつれどのように変化しているかをFACT-Graphは可視化する。この可視化情報は図１に示すような共起グラフとして表示され、分析者はこの画像からトレンドについて考察する。

一方、昨今のテキストマイニングでは、人間を中心として分析をしていく方法論が注目を浴びている[8]。そのテキストマイニングでは、知識獲得プロセスにおける分析者とツールとの相互作用性を強調した対話型システムとして実現される。具体的に、テキストマイニングの結果から分析者が考察し、その考察内容をテキストマイニングに反映させ、再び考察するといったことを繰り返していく。この対話型テキストマイニングを効率的に進めるために、分析者の支援を行うシステム（以降、テキストマイニングシステム）が望まれている。そして、そのシステム上で分析に必要な情報ソースやエビデンスの参照、ユーザの意志反映のサポートなどを行うことが求められている。

本論文では、FACT-Graphを対話型テキストマイニングシステムとして発展させることを目的に、FACT-Graphの分析を支援するシステムを開発する。このシステムを本論文ではLoopoと呼ぶ。Loopoは分析者がFACT-Graphの考察をしやすいように、情報ソースの参照やパラメータの変更を可能にし、またFACT-Graph自身を操作するインターフェイスを提供する。

本論文は以下の構成を取る。２章ではトレンド分析やテキストマイニングシステムについての関連技術について述べ、その中で本論文での中心となるFACT-Graphを紹介する。３章では２章での考察をもとに本論文で提案するシステムであるLoopoについて述べる。４章ではLoopoによって日本の犯罪データのトレンド分析について述べ、最後に５章にて本論をまとめる。

† 神奈川工科大学　Kanagawa Institute of Technology
‡ 大阪府立大学　Osaka Prefecture University

2．関連研究

2.1　トレンド分析

トレンド分析によって得られるものは、期間において何がどのように変わったかという質問に関する答えである。この答えを得るために、重要と思われる語句をキーワードとして用いたトレンド分析では、語句とその関連全体を把握し考察する必要がある。

Montres-G-Gomezらは固定された2期間において確率分布の差分から求める関係を定義している[9]。その関係はephemeral associationと呼ばれ、語句感の関係を明らかにすることでトレンドの抽出を目的としている。しかし、この手法は関係の抽出に伴うトレンドの抽出が目的であり、可視化による大域的なトレンド把握が目的ではない。一方、FeldmanらのTrendGraph[5]やHavreらのThemeRiver[6]はキーワードの可視化を目的として提案されている。ThemeRiverはキーワードの連続したトレンドを川の流れのように表現し、TrendGraphはある2期間における関係の変化によりトレンドを抽出しようとしている。しかしながら、ThemeRiverは語句関係については考慮しておらず、またTrendGraphは関係のトレンドに注目しており語句自体のトレンドを考慮していない。

2.2　FACT-Graph

2.1の背景のもと、我々は新たにFACT-Graphと呼ぶ可視化手法を提案している[7][10]。FACT-Graphは2期間における語句自身の変化と語句間の関係の変化を共に可視

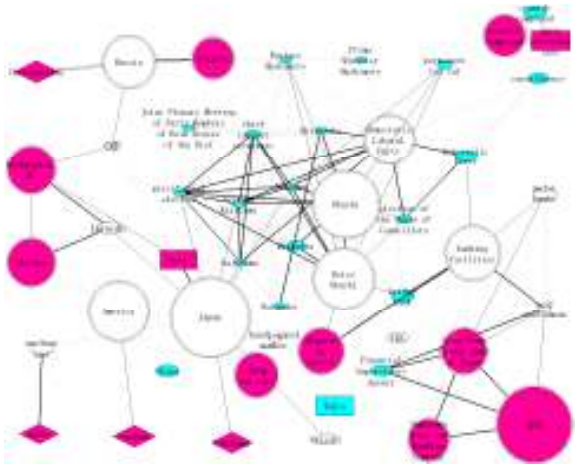

図1. FACT-Graph[10]

化している。この FACT-Graph と 2.1 で述べた手法との比較を表 1 に示す。

FACT-Graph は頂点（以下、ノード）と枝（以下、リンク）の集合によって表現されており、キーワードの変化や語句間の関係はそれらの中に表されている。このキーワードの変化を求めるために、FACT-Graph ではクラス遷クラス遷移分析を用いている[11]。

クラス遷移分析は、各期間において、Term Frequency（TF）と Document Frequency（DF）から語句を大きく A, B, C, D の 4 つのクラスに分割する。このとき、Class A は TF、DF 共に高いものが、Class B は TF が高く DF が低いものが、Class C は TF が低く DF が高いものが、Class D は TF も DF も低いものが割り当てられる。そして、ある 2 期間の間で語句が所属しているクラスがどのように変化しているかによってトレンドを表現する。その結果は表 2 のように意味づけされている。例えば、もしある期間において Class A に属する語句が次の期間において Class D に遷移したとき、その語句は”Fadeout”したと見なされる。FACT-Graph ではこのクラス遷移分析の結果を色によって認識する。具体的に、盛んに成ってきている語句は赤、変わらないものは白、また話題から消えかけているものは青で表している。

また、FACT-Graph はキーワード間の関係を表現する。この関係として共起関係を用いており、この関係を用いることで重要ではないと見なされている単語から、これから発達しそうなキーとなる単語を見いだす。さらに、クラス遷移分析の結果と、共起関係から形成されるクリークから暗黙的に話題にされているトピックなどを分析することができる。

この FACT-Graph を生成するための手順は以下のステップからなる。

1. キーワードを分析期間に従って分割する。例えば、10 月 1 日から 10 月 31 日までを分析しようと考えた時、10 月 1 日から 10 月 15 日までのデータと 10 月 16 日から 10 月 31 日までのデータに分割する。
2. 形態素解析により語句を抽出し、TF-IDF アルゴリズムによりキーワードを抽出する[1]。ここで、用いる TF-IDF アルゴリズムは Harrnan[2]による方法を採用する
3. クラス遷移分析と共起関係を抽出する。
4. Step 3 で得た結果を可視化する。その可視化には、Graphviz を用いている[12]。

2.2.1 FACT-Graph 生成におけるパラメータ

FACT-Graph は時間、語句、共起関係という 3 つの要素からなる。故に、FACT-Graph の生成パラメータとして、その 3 つの要素に関連したものを設定する必要がある。例えば、採用すべきキーワードの TF, DF の閾値、分析期間、共起関係の種類や共起関係の閾値などがある。また、その他にも FACT-Graph 上に表すキーワードの最大数など設定する。FACT-Graph の分析者は、これらのパラメータを設定することで FACT-Graph を生成する。

2.3 テキストマイニングシステムへの発展

知識発見の場面において、対話型のシステムは数多くある。我々の知る限りでは、1990 年代において人間の主観情報を活用した知識発見のコンセプトが紹介され、テキストマイニングシステムは、そのコンセプトを元に提案されている[8]。このようなシステムでは、分析対象を容易に出力可能であるといった点、そしてインタラクティブ性が重要視される。例えば、昨今では Polaris[13]と呼ばれるシステムが提案されている。Polaris は KeyGraph[14]におけるパラメータ設定などを容易にし、分析を促進するために開発されている。

表 1　関連研究

	Visualization	Relationships Between topics	Keywords Trends
Ephemeral association		✓	✓
ThemeRiver	✓	✓	
Trend Graph	✓	✓	
FACT-Graph	✓	✓	✓

表 2　クラス遷移分析

		After			
		A	B	C	D
Before	A	Hot	Cooling	Bipolar	Fade
	B	Common	Universal	-	Fade
	C	Broaden	-	Locally Active	Fade
	D	New	Widely New	Locally New	Negligible

2.3.1 FACT-Graph の問題点

従来の FACT-Graph を対話型テキストマイニングシステムとして発展させるために、我々は FACT-Graph とテキストマイニングシステムによる分析方法論を比較検討した。そのとき、改善すべき課題がいくらかあがった。

その課題の一つとして、分析者の気づきを反映しづらいといった問題点がある。FACT-Graph から得た考察は、新たなキーワードの選定やパラメータ値の調整といった形で反映される。しかし、FACT-Graph にあらたなキーワードを含めるためにはパラメータの値を設定することで調整していたが、特定のキーワードのみを反映するのは難しい。これはトレンド分析によって得られる知識発見プロセスを妨げる要因になりうる。

さらに、情報の参照性に関しても問題がある。語句の重要性や関係は分析期間により変化し、その変化を知るには元となる情報を参照する必要がある。しかしながら FACT-Graph は画像であるため、全体のトレンドを把握することはできるが情報ソースへのリンクを保持しないため、情報を直接参照することができない。これは分析者がインタラクティブに FACT-Graph 自身を操作できないことであり、同様に知識発見プロセスを妨げる。

2.3.2 FACT-Graph の拡張要件

これらのことから本論文では FACT-Graph を対話型テキストマイニングシステムとして拡張するために、以下の要件でシステムを作成した。

- FACT-Graph の基本的な要素はすべて満たしている

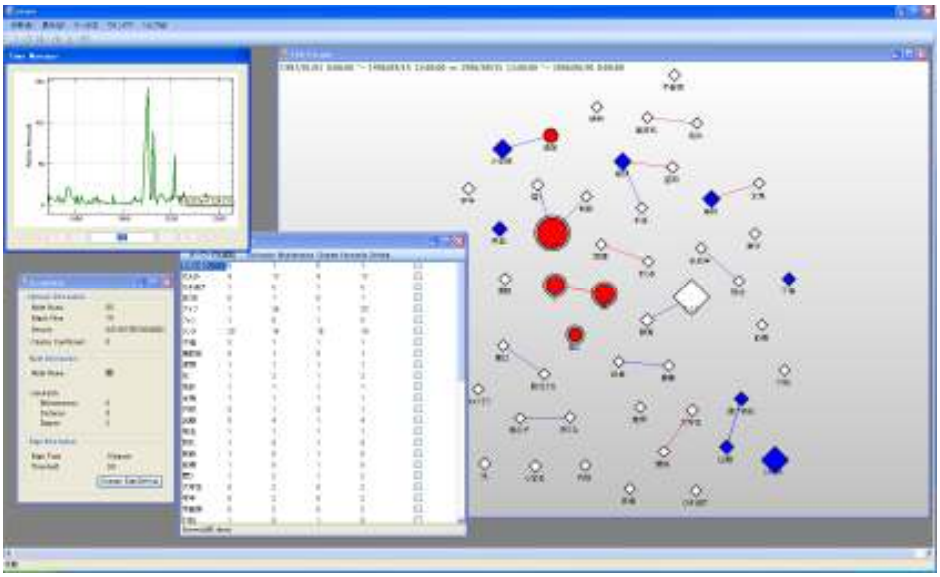

図 2. Loopo スクリーンショット

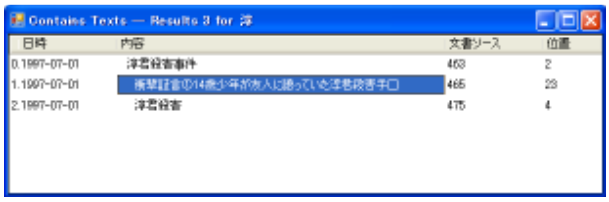

図 3 KWIC 機能

- 分析者の考察結果の反映をシステム上で可能にする。ここではその反映結果を新たな語句の追加、パラメータの変更に限定する。
- GUI により FACT-Graph 自身をインタラクティブに操作可能にする。この操作は FACT-Graph に表示されている要素に対して、移動や変更、動作の固定を提供する。
- FACT-Graph からキーワードの情報ソースを直接参照可能にする。

これらの要件を元に開発したシステムを我々は Loopo と呼ぶ。

3 . Loopo

FACT-Graph を対話型なシステムとして拡張するためには、FACT-Graph を用いた分析プロセスにおいて、ユーザのインプットとアウトプットが繰り返して行われるようにすることが望ましい。Loopo は FACT-Graph による分析プロセスをそのように改善し、支援するものである。Loopo は FACT-Graph を生成する他、その生成に必要なパラメータの設定、また情報ソースへの参照を容易にする。

図 2 はあるテキストデータを Loopo に入力し、FACT-Graph を生成した例である。Loopo は FACT-Graph View, Time Manager, Keyword Manager, Graph Info と呼ばれる 4 つの基本となるウインドウからなる。分析期間やキーワードに関するパラメータは、これらのウインドウから設定可能であり、そしてそのパラメータは複数のウインドウで共有される。

Loopo による分析は時系列テキストデータの入力から始まる。データ入力の後、Loopo は FACT-Graph の生成プロセス同様に分析期間によりテキストデータを分割し、分割した各データから形態素解析と TF-IDF により語句を抽出し、クラス遷移分析や共起関係を求め可視化する。

3.1. FACT-Graph View

FACT-Graph View はテキストデータから生成された FACT-Graph を表示する場所である。FACT-Graph View では FACT-Graph をリアルタイムでバネモデルを元に配置を行っていく。一方、分析者はこのウインドウにある FACT-Graph のノードを移動し、固定化させ、また削除することが可能である。特にノードの固定化機能によって気になるノードをある位置にとどめることができ、複数期間に渡ってそのノードの指すキーワードがどのように変化するかを確認できる。また、FACT-Graph View では気になるキーワードから情報ソースを参照することができる KWIC 機能を提供している（図 3）。これにより、従来では情報源を探し出すなど手間を省くことが

3.2. Time Manager

時系列データを用いた分析においてどの期間を分析するかを把握し、設定することは重要なことである。一般的に文書量の変化を把握することは一つの分析の手掛かりになる。Time Manager は、時系列におけるテキストデータの量の推移を線グラフで表し、また分析期間の設定を提供する。また、Time Manager では複数の連続した分析期間を順番に推移させ、その時に応じた FACT-Graph の生成を促す役割を持つ。

3.3. Keyword Manager

Keyword Manager は FACT-Graph に表示されている語句の一覧を表示する。その情報として、2 期間における TF と DF の値、また TFIDF によるウエイト値など語句に関する情報を提示する。また、このウインドウを経由してユーザはキーワードの追加や削除、また辞書のメンテナン

スを行うことができる。また、TF や DF の閾値などキーワードに関するパラメータの他、形態素解析で用いる辞書を設定することもできる。

3.4. Graph Info

FACT-Graph を分析する際の一つの要素として、密度やクラスタ係数といったネットワーク情報がある。ネットワーク情報はそれを用いたキーワード抽出なども開発されているように、キーワードを用いた共起グラフにおいて重要な指標となる[15]。GraphInfo は現在の FACT-Graph のネットワーク情報を表示する。また、各ノードに対して媒介中心性や近接中心性を計算し、その情報を閲覧することが可能である[16]。その他、Jaccard 係数や Simpson 係数といった共起係数のタイプや、関係があると見なす閾値の値の設定なども GraphInfo から可能である。

3.5.入力データ

今回用いた入力データは改行コードなどを正しく認識するためにも XML 形式を用いている。その XML 構造を図 4 に示す。この XML はルート要素として DOCS を持ち、各文章は DOC 要素内に記述される。DOC 要素は DOCNO、DateTime、TEXT 要素を持ち、それぞれ文章の ID、文章の時間、文章本文を記述する。

3.6. FACT-Graph の意味

Loopo における FACT-Graph では各ノードが色と形で表される（図 5）。ノードの大きさは分割した期間（前期と後期）における TF の大きさのうち、大きい値を表現している。各リンクは前期と後期での共起関係の有無を表しており、前期と後期において共にリンクがある場合は黒色、前期になく後期になり出現したリンクは赤、逆に後期になり消滅したリンクは青で表している。すなわち、ノードとリンクともに新たに現れてきているものは赤で、消滅してきているものは青で示される。また、各ノードは後期におけるキーワードのクラスを表している。また黄色の四角は新たにユーザが気づきとして追加したものとしている（図 6）。

4． Case Study

4.1.環境

本章ではケーススタディとして、1987 年から 2007 年における朝日新聞の記事を対象に、Loopo によって分析した結果を報告する。今回のケーススタディでは、特に上記期間の犯罪記事に関する 2971 件の見出しを用いた。これに関するデータの詳細を表 1 に示す。また、今回用いたパラメータとして、各期における TF-IDF ウエイトの大きいキーワードを上位 30 個抽出し、共起係数の種類として Simpson 係数を用いた。また、見出しだけの関係とデータ量が本文よりも少ないことから、採用する語句の TF, DF、共起係数の閾値が 0 以上のものを分析対象とした。

4.2.分析と考察

まず、最初にデータセット全体を二つに分割し、それから FACT-Graph を生成した。その結果を図 6 に示す。今回は特に盛んになってきているキーワード、つまり赤いノードに注目してみると、この図から、『淳』といった

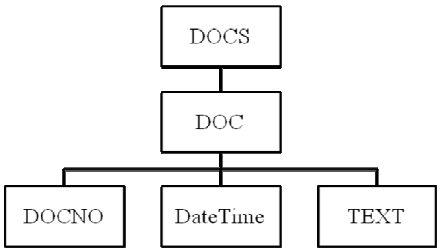

図 4 入力データの XML 構造

表 3 分析データとパラメータ情報

The number of articles		2971
Analysis Span		1987-2007
Co-occurrence Type		Simpson
Parameters	Keywords	30
	Thresholds(TF, DF)	0
	Thresholds (Co-occurence)	0

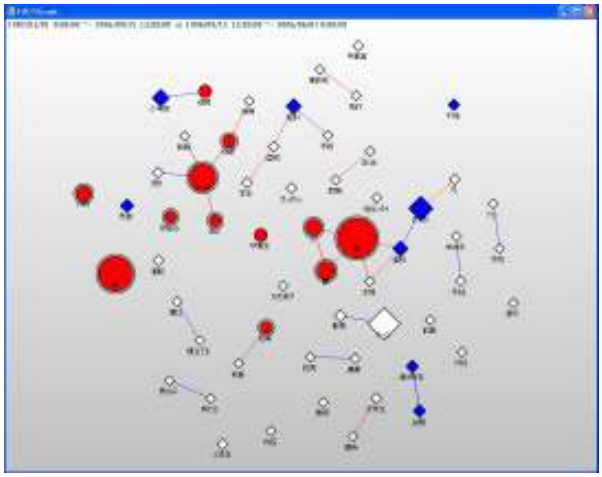

図 5 FACT-Graph の結果
（1987 年から 2007 年の新聞記事のタイトル）

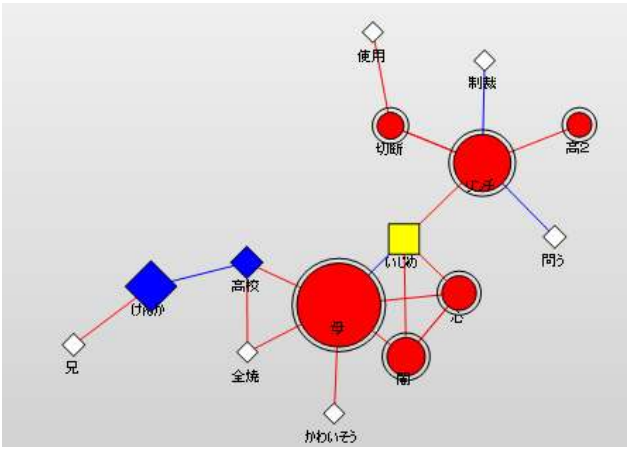

図 6 『いじめ』を追加した結果
（1987 年から 2007 年の新聞記事のタイトル）

固有名詞の他、『母』、『ナイフ』、『心』、『闇』と

いったキーワードがこの期間において盛んになってきていることがわかる。

このとき、『心』や『闇』と言ったキーワードは KWIC 機能から『心の闇』という一つのフレーズで用いられることが多く、ほとんど以前は登場していなかったこともわかった。そして、この 2 つのキーワードについて詳細をみていくと、『いじめ』というキーワードがいくらか見られるようになってきた。

ここで、この『いじめ』を新たな気づきとして入力する。この追加した結果を図 7 に示す。その結果、今まで今まで分割されていた新たな島との接続となって現れた。また、いじめのキーワードの媒介中心性は他のものと比べて大きいこともわかる。このことから、『いじめ』というキーワードはこの 20 年における犯罪において重要なキーワードである可能性が高いことがわかる。

また、その他のキーワードについて気になるワードとして少年がある。その少年を Loopo によって追加した FACT-Graph を図 8 に示す。この結果からわかるように、少年は多くのキーワードと関係をもつ。そのため、一つの重要なキーワードとも見られるが、各記事に遍く存在することから情報量が少ないとみなされ、TF-IDF によって採用している FACT-Graph では削除されていることが推測できる。また、図 8 において少年は数多くのものと関係を持つために一つのクラスタを形成し、密集しがちである。そのため実際はノードが把握できないが、Loopo でノードの固定化などの操作を行うことでノードを把握しやすい位置に表示することができる。

続いて、Time Manager から犯罪記事数の変化を見たとき、1996 年後半から 1997 年前半と 1997 年後半から 1998 年前半の間に通常時と比べて記事数が非常に多くなっていることがわかる（図 9）。そこで、まず Loopo によって前者の期間を分析期間とし前期と後期の 2 期間に分け出力した。この結果を図 10 に示す。図 10 では、全体の期間で抽出された『淳』というキーワードが盛んになっている事がわかり、このキーワードで調べた結果、神戸における中学生の少年犯罪について表されていることがわかった。続いて、後者の期間を分析した。その結果、この時期においては大量のキーワードの推移が確認できた(図 10 左図)。しかしながら、重要ではないと思われる単語なども大量に出力され、分析が困難であった。このとき、Loopo を用いることで、情報を参照しながら不要と思われる単語を消去していき、代表的と思われる単語のみを残すことができた(図 10 右図)。その結果、後者では、前者の事件に関しての関連報道についての内容や、新たな少年犯罪についての報道が為されていることがわかった。

4.3.考察と評価

今回のケーススタディでは、Loopo を用いて分析を行った。その際、気づきを即座に反映するために新たなキーワードを追加することをスムーズにできた。また、混雑した情報からも内容を確認しながら不要な単語を消去していくことでより分析しやすい FACT-Graph の出力を実現できた。しかしながら以下のような考慮すべき点もまた浮上した。

- 複数人での協働環境

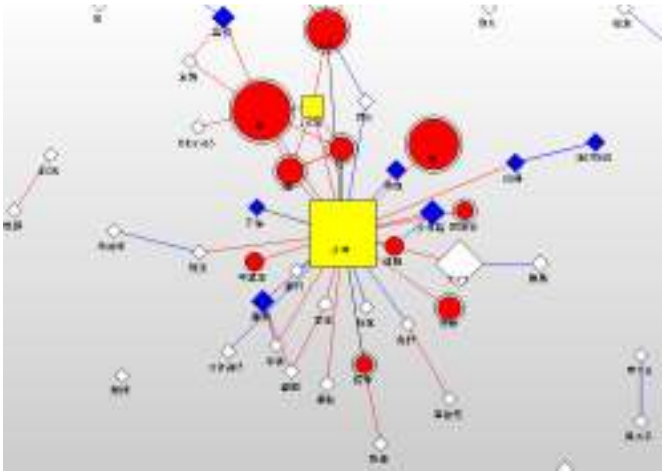

図 7　『少年』を追加した FACT-Graph

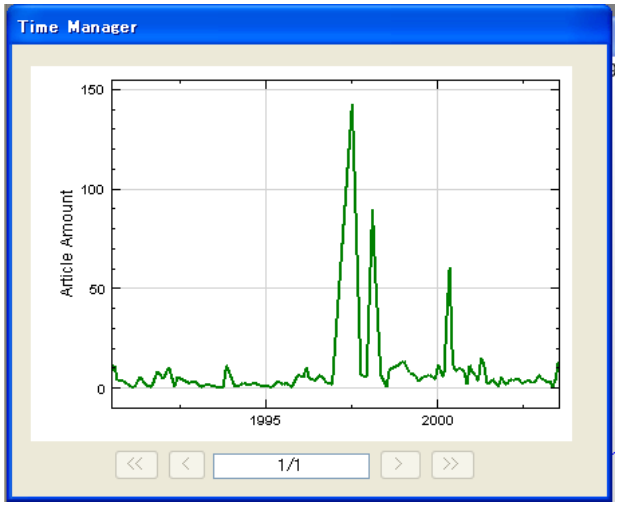

図 8　Time Manager からわかる文書数の変遷

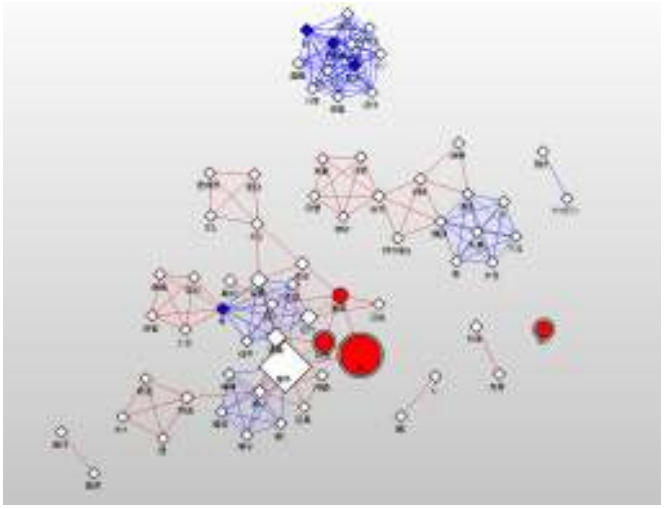

図 9　FACT-Graph の結果
（1996 年 10 月～1997 年 11 月の新聞記事タイトル）

Loopo はトレンド分析を主に扱うものであるが、複数人によって分析され、またその結果を共有することも考え

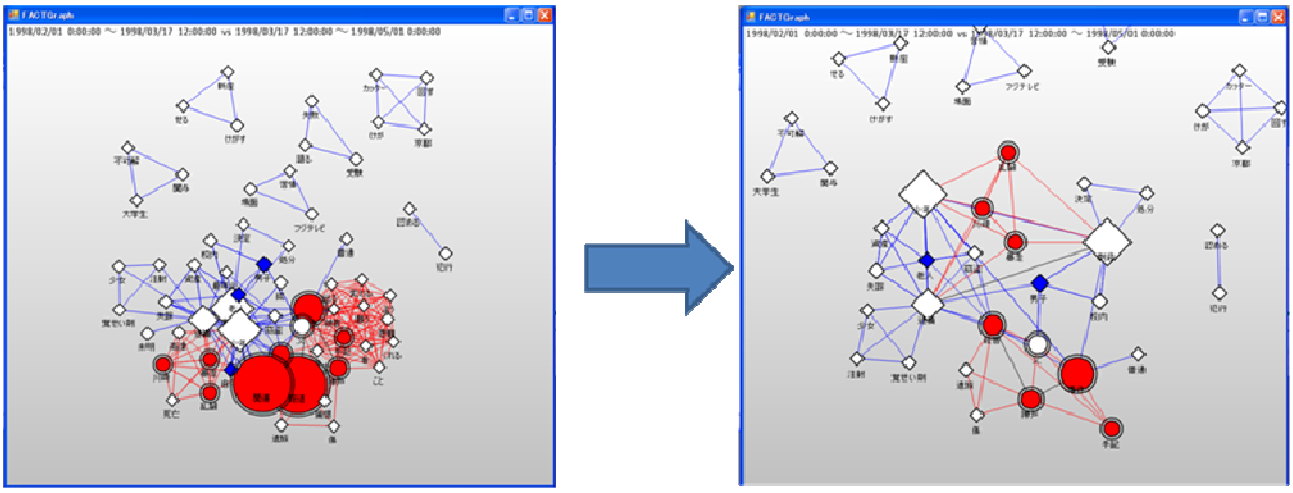

図 10　Loopo による FACT-Graph 上のキーワード選定

られる。しかしながら、Loopo はその点を考慮しておらず、アノテーション機能やメモ機能など複数人で結果を共有する機能を実装する必要がある。

● 分析のプロセスについて

従来と比べて Loopo を使うことでパラメータの変更などスムーズに分析ができた。しかしながら、事前の知識があるために内容が理解できる部分もあり、事前知識や経験なしで分析することは難しい。どのような人でもより容易に分析できるようなプロセスなどを開発する必要がある。

5．おわりに

本論文では Loopo と呼ばれる FACT-Graph の分析を支援するシステムを提案した。FACT-Graph を分析するときには、1）分析者の考察した結果を分析に反映させにくい、2)画像として結果が提供されるため、情報探索のための操作が失われているといった問題点があった。そこで、Loopo では a)分析者の考察結果を新たな語句として追加する機能やパラメータとして反映させる機能、また、b) FACT-Graph 自身を操作可能にする機能を付け加え、FACT-Graph による分析を容易にできるよう可能にした。Loopo は FACT-Graph 自身の操作の他、分析に必要な情報を提供する。そして、Loopo の有用性を確認するため、犯罪に関する記事を Loopo により分析し、Loopo の有用性を検証した。本システムを元に、購買履歴からこれから売れる商品のトレンド予測など、FACT-Graph の利用についてより検証をしていく予定である。

謝辞

本研究を実施するに当たり、データ収集と提供に協力してくださった神奈川工科大学 鷹野孝典先生、大木俊氏に感謝する。また、本研究は科研費（21760305）の助成を受けたものである。

参考文献

[1] Salton, G.: “Automatic Text Processing”, Addison-Wesley Publishing Company (1989)

[2] Harman, D.: “Ranking algorithms”, in Information Retrieval, chapter 14. Prentice Hall (1992).

[3] 奥村 学, 難波 英嗣, “テキスト自動要約”, オーム社 (2005)

[4] Yamanishi, K., Li, H.: “Mining Open Answers in Questionnaire Data”, IEEE Intelligent Systems, Vol. 17, Issue 5, pp. 58-64 (2002)

[5] Feldman, R., Aumann, Y., Zilberstein, A., and Ben-Yehuda Y.: “Trend graphs: Visualizing the evolution of concept relationships in large document collections”, Second European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD 1998), pp. 38-46 (1998)

[6] Havre, S., Hetzler, B., and Nowell, L.: “ThemeRiver(TM): In search of trends, patterns, relationships”, IEEE Trans Visualization Computer Graphics 8, pp. 9-20 (2002)

[7] 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”, 電学論 C, Vol.129, No.3, pp.545-552(2009).

[8] Brachman, R. and Anand, T, “The Process ofKnowledge Discovery in Databases: A Human CenteredApproach”, A KDDM, AAAI/MIT Press, pp. 37-58 (1996)

[9] Montes-y-Gomez, M., Gelbukh, A., and Lopez-Lopez, A.: “Mining the News: Trends, Associations and Deviations.” Computacion y Sistemas 5(1): pp. 14-25.

[10] Saga, R., Terachi, M., Sheng, Z., and Tsuji, H.: “FACT-Graph: Trend Visualization by Frequency and Co-occurrence”, in Lecture Notes on Artificial Intelligence (LNAI 5243: Ed by A. Dengel et al.(Eds)), Springer-Verlag Berlin Heidelberg, pp.308-315 (2008)

[11] Terachi, M., Saga, R., and Tsuji, H.: “Trends Recognition in Journal Papers by Text Mining”, Proceedings of IEEE International Conference on Systems, Man & Cybernetics (IEEE/SMC 2006), pp. 4784-4789 (2006)

[12] Ellson, J., Gansner, E. R., Koutsofios, E., North, S. C., and Woodhull, G.: Graphviz - open source graph drawing tools, Graph Drawing, pp. 483-484 (2001)

[13] Okazaki, N. and Ohsawa, Y.: “Polaris: An Integrated Data Miner for Chance Discovery”, Proceedings of Workshop of Chance Discovery and Its Management (in conjunction with International Human Conputer Interaction Conference (HCI2003)), Crete, Greece (2003).

[14] Ohsawa, Y., Benson, N. E., and Yachida, M.: “KeyGraph : Automatic Indexing by Segmenting and Unifing Co-occurrence Graphs”, IEICE D-Ⅰ, Vol. J82-D-Ⅰ, No. 2, pp. 391-400 (1999)

[15] 松尾 豊，大澤 幸生，石塚 満, ”Small World 構造を用いた文書からのキーワード抽出”，情報処理学会論文誌，Vol.43, No.6, pp. 1825-1833 (2002)

[16] Freeman, L. C.: “Centrality in social networks: Conceptual clarification”, Social Networks, Vol. 1, No. 3, pp. 215-239 (1979).