人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第13回)
SIG-AM-13-01

時系列データの探索的分析を支援する可視化システム: 記事と時系列データのアラインメント方式の提案

Visualization System for Exploratory Analysis of Time-series Data: Alignment Method between Articles and Time-series Data

内藤　峻[1*]　古田　遼樹[2]　松下　光範[3]
Shun Naito[1]　Ryoki Furuta[2]　Mitsunori Matsushita[3]

[1] 関西大学大学院　総合情報学研究科
[1] Graduate School of Informatics, Kansai University
[2] 数研出版株式会社
[2] Suken Shuppan
[3] 関西大学　総合情報学部
[3] Faculty of Informatics, Kansai University

Abstract: The goal of our study is to support a user's analysis of time-series data in an exploratory manner. Such exploratory analysis requires repeated access to various types of information related to the user's interests such as texts and numerical data. To support such the user's analysis, we have proposed a system that visualizes temporal changes in time-series data and presents the causes of those changes with the data. In this paper, we improve the system by adding an alignment function between news articles and time-series data. By using this function, the user can find articles that relates to the time-series data easily.

1　はじめに

時系列データとは，熊本地震の被害者数や気温など時間の経過に伴って変化するデータである．このような時系列データは意思決定や問題解決の場面で役立てられている [1]．意思決定や問題解決の場面では，時系列データの値の変化やその変化の要因を分析することで，有益な情報や新たな知見を得ることが重要である．しかし，このような時系列データの分析は仮説の生成や検証を探索的に繰り返す負荷の高い作業であるため，ユーザがこのような探索行為を円滑に行うことが難しいという問題がある．そこで本研究では，ユーザの興味や関心に応じて様々なモダリティの情報へのアクセスを繰り返しつつ時系列データを分析するための支援システムの実現を目指している．その端緒として，著者らはこれまでに，新聞記事と地図，統計データを対象に，ユーザが時系列データの経時的変化とその変化の要因を把握できるようにする可視化インタフェースを提案してきた [2][3]．本稿では，そのインタフェースに組み込む機能の 1 つとして，新聞記事と時系列データのアラインメント方式を提案する．この方式をシステムに組み込むことにより，効率的に時系列データと文章を対応付けることができる．さらに，新たに実装したインタフェースの機能について述べる．

2　システムの全体像とこれまでの取り組み

図 1 に本システムの目指す構成を示す．現状のシステムは新聞記事 DB，統計 DB，地図 DB を人手で作成している．人手での作成は，効率性や網羅性，リアルタイム性の点で問題がある．そのため，システムに用いるデータは WEB から抽出することを考えている．新聞記事 DB には，クローラを用いてあるトピックに関する記事を収集し，スクレイピング技術を用いて本文や見出しを抽出する．また，抽出された見出しや本文から自然言語処理技術を用いて日付や国名，出来事に関する文を抽出することを考えている．統計 DB や地図 DB は，オープンデータとの連携を検討している．

*連絡先：関西大学大学院総合情報学研究科知識情報学専攻
〒 569-1095 大阪府高槻市霊仙寺町 2-1-1
E-mail: mat@res.kutc.kansai-u.ac.jp

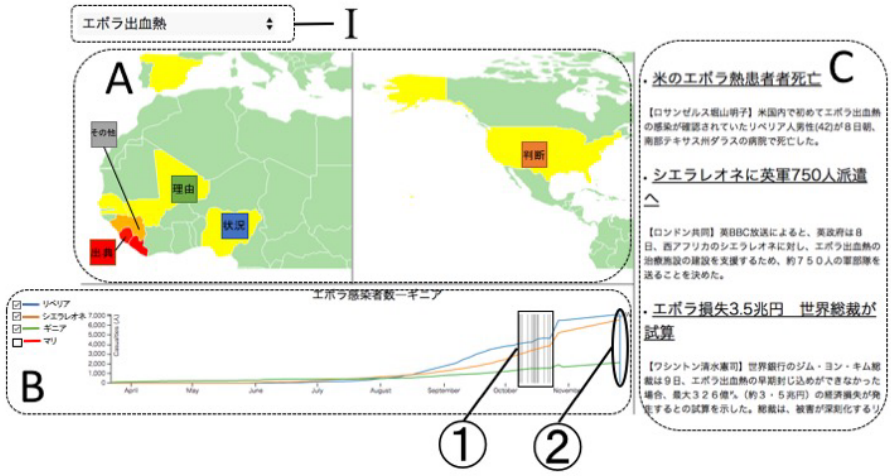

図 2: システムの全体像

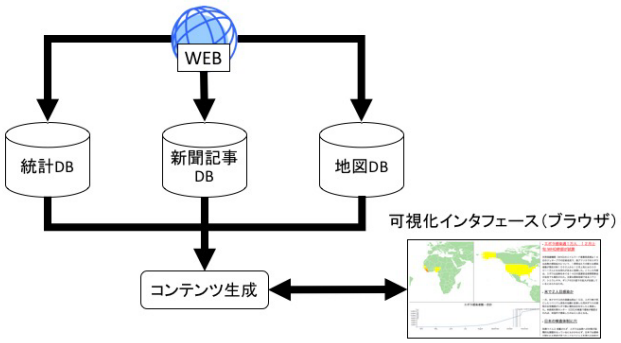

図 1: システムの構成

2.1　システムの全体像

著者らは，ユーザが効率的に時系列データを分析することができる可視化システムについて研究している．図 2 に理想のシステムの全体像を示す．システムは，トピックを選択する選択ボックス (図 2–I)，地図を表示する地図ペイン (図 2–A) とグラフを表示するグラフペイン (図 2–B)，記事を表示する記事ペイン (図 2–C) で構成されている．選択ボックスには,「エボラ出血熱」や「台風」といったトピックがプルダウン形式で表示される．ユーザがトピックを選択すると，そのトピックに関連する統計量がグラフとしてグラフペインに描画され，同時に地図ペインには関連する地図を，記事ペインには関連する記事が表示される．また，地図 (図 2–A) に記事の有無を表すアノテーションとしてアイコンが付与されている．地図上のアイコンは文献 [4] を参考に，5 種類 (理由，背景，状況，出典，その他) を考えている．ユーザがアイコンをクリックすると，そのアイコンに対応する記事が表示される．さらに，グラフペイン (図 2–B) には，複数の統計グラフと凡例が提示されるようになっている．ユーザは凡例の左にあるチェックボックスに比較したい国を選択することでグラフペインに複数の統計グラフを描画することができる．これにより，ユーザは興味を持った国同士の統計量の変化を比較することで他国からの影響や規模を詳しく知ることができる．

現状のシステムには，トピックを選択する機能や地図上のアイコン，複数の統計グラフを描画する機能は実装されていないが，将来的にはこれらの機能を全て実装するつもりである．

2.2　これまでの取り組み

図 3 に，本研究で対象とするデータの関連性を示す．システムの機能は，図 3 に示される情報アクセス行為を行えるように実装された．グラフペインには，ユーザによって選択された日付を表示する青い線 (図 2-①) と記事の有無を表すアノテーション (図 2-②) が表示されている．ユーザは青い線を左右にドラッグして日付を選択することができる．日付を選択すると，その時点の統計量が地図にマッピングされる．これによって，ユーザはグラフを見て興味を持った時点の統計量の変化が地理的な影響を受けいているのか，周辺の国に影響を与えているのか，把握することができる (図 3-①)．また，青い線をアノテーションに重ねると，その時点で起こった出来事について記述された記事がハイライトされる．これによって，ユーザはグラフを見て興味を持った時点の統計量の変化の理由や背景を知ることができる (図 3-⑥)．さらに，地図にマッピングされた国はクリックすることができる．国をクリックすると，

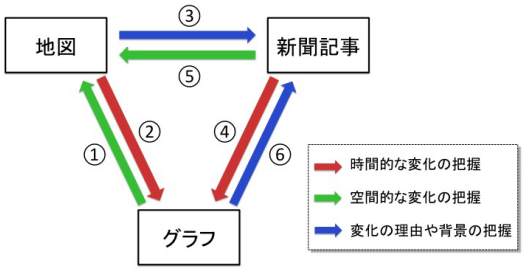

図 3: 対象とするデータの関係性

その国の統計量がグラフとしてグラフペインに描画される．これによって，ユーザは興味を持った国の統計量の変化を詳しく見ることができる (図 3-②)．加えて，記事をクリックすると，記事に含まれる日付がグラフ上に表示される．これによって，ユーザは記事を見て興味を持った出来事がグラフのどの時点で起こったのか把握することができる (図 3-④)．

残りの新聞記事から地図へのアクセスと地図から新聞記事へのアクセスを行うユーザの振る舞いと機能 (図 3-③と図 3-⑤) については，3 節で述べる．

現状のシステムでは，記事の有無を表すアノテーション (図 2-②) は，人手で記事の本文を見て日付を抽出し，グラフの日付と対応付けられている．人手での対応付けは効率性や網羅性，リアルタイム性に欠けるという問題がある．4 節では，この問題を解決するために，記事の有無を表すアノテーションの付与の自動化について検討した方式について報告する．

3 新聞記事と地図との間のインタラクション

3.1 デザイン指針

本節では，先行研究で検討した新聞記事と統計データ，地図の特徴とデザイン指針について述べる [3]．

新聞記事は，ある時期における出来事やその出来事が起こった原因，場所，統計量，その統計量の具体的な値への言及や予測，記者の意見などが書かれている．そのため，出来事が起こった理由や背景を理解する上で有用である．しかし，新聞記事に書かれている統計量の値は近似値が用いられているため正確でなかったり，記者の観点で纏められていたりするため，客観性に欠ける．

統計データは，ある観測された場所において，ある時点の事象について測定された値である．例えば，人口統計や外国為替相場の推移などがこれに当たる．これらの統計データの値は，政府や国際連合の専門機関などが実施している厳密な環境において観測されていたり，センサを用いて取得される．そのため，これらのデータは正確である．

地図は地球や地表，架空の世界の全部もしくは一部を平面上に縮尺表現したものである．例えば，地球全体もしくは大部分を表現している世界地図や統計データを地図上に表した統計地図などがこれに当たる．地図は空間的な位置関係や方向，距離，面積，形，高さを知る上で有用である．また，ある時点の統計量を地図にマッピングすることで，出来事の規模や地理的な広がりを把握できるといった特徴を持っている．

これらの情報はそれぞれ単体でも用いることができるが，ユーザの興味となる要素をトリガとしてインタラクティブに情報を提示することで，円滑な情報アクセスが可能なインタフェースを実現できると考えている．

図 3 で示したデータの関係性から，ユーザは，統計地図を見て出来事の地理的な影響や規模を把握しつつ，新たに統計値がマッピングされた国や統計量の特徴的な変化に興味を持ち，その理由や背景を知るために新聞記事を参照する (図 3-③)．また，新聞記事で言及されている出来事に興味を持ち，その出来事の地理的な影響や規模を把握するために統計地図を参照する (図 3-⑤)．本研究では，このような情報アクセスを行えるように，システムの機能をデザインした．これによって，ユーザの探索行為を円滑にする．

3.2 実装

本システムでは，対象データとして 2014 年の西アフリカエボラ出血熱流行に関する統計データ [1] とそれに言及している新聞記事データを用いた [2]．統計データは，2014 年 3 月 22 日から 2014 年 11 月 26 日までの日付と累計患者数，各国の感染者数を csv 形式のデータとして纏めたものである．新聞記事データには，2014 年 10 月 9 日から 2014 年 10 月 30 日までの毎日新聞の記事 (計 18 記事) を用いた．次に，これらの記事からエボラ出血熱について書かれた記事を選び，その記事から出来事に関する文のみを抜き出した．さらに，その文から出来事が起こった日付と場所を示す国名を抜き出し，抜き出した出来事に関する文章の統計量名，を付与したものを csv 形式のデータとして用意した．日付は記事に含まれている「数字＋日」を抽出し，年と月を加え「年/月/日」とデータを正規化した．出来事に関する文は，「米国内で初めてエボラ出血熱の感染が確認されていたリベリア人男性 (42) が 8 日朝，南部テキサス州ダラスの病院で死亡した」といった文の一段落分を抽出した．国名は，本文中の「米国」や「リベ

[1] http://ja.wikipedia.org/wiki/2014 年の西アフリカエボラ出血熱流行

[2] 現在は，統計量名の種類としてエボラ感染者数のみしか対応していない

リア」等の名詞を抽出し,「米国」や「米」,「アメリカ」など同じ国のことを指し示している表現は「アメリカ」といったように1つの単語に統一した．統計量名は，本文の「死亡した」や「死者」といった語句から「エボラ死者数」,「感染が確認された」や「感染者数」といった語句から「エボラ感染者数」を人手で判断し，統計量の名称を付与した．また，統計量名がない場合は「非統計情報」とした．記事ペインの各記事のスニペットには，これらの情報がタグ付けられている．

3.1節で述べたデザイン指針に基づき実装した機能について述べる．ユーザが国をクリックすると，その国名が含まれる記事がハイライトされる機能を実装した．システムは，国がクリックされたことを判断すると，選択された国名を記事モジュールへと引き渡す．記事モジュールは引き渡された国がタグ付けられている記事をハイライトする．これにより，ユーザは地図を見て興味を持った国に関する統計量の変化の理由や背景を把握することができる．

記事をクリックすると，その記事に含まれる国がハイライトされる．システムは，記事がクリックされたことを判断すると，データがタグ付けられた記事のスニペットから記事中に含まれる国を地図モジュールへと引き渡す．地図モジュールは引き渡された国をハイライトする．これにより，ユーザは興味を持った記事の出来事の影響や規模を把握することができる．

4 新聞記事と時系列データのアラインメント方式

4.1 対象とする課題

現状のシステムでは，時系列データと文書との対応付けは人手に委ねられており，効率性や網羅性の点で問題がある．この問題を解決するため，本研究では新聞記事やブログなど時系列データの変化の理由や背景が記述された文書が，時系列データのどの期間に対応しているのかを自動的に推定し，それらを紐付けて提示する手法の実現を目指す．

時系列データと文書を対応付けるための研究は多数行われている．例えば，小林らはグラフで示された数値情報を自然言語テキストで説明する手法を提案している [5]．提案手法は選択体系機能言語理論を用いてグラフの特徴と言語表現の関係を分析し，それに基づきテキストを生成している．また，日経平均株価のグラフとその動向を説明するテキストを用いてグラフとテキストが協調的に提示される手法を提案している [6]．この手法では，長期的な動向には，グラフの表示状態に合わせてテキストを要約し，提示している．短期的な動向には，人間が視覚的に捉えるグラフの挙動を説明するのに適切なテキストが生成される．

また，Ahmad らや Boyd は時系列データに Wavelet 解析を行い，グラフの特徴を特定し，自然言語テキストを生成する手法 [7][8] を，馬野らは全体的傾向と局所的特徴を組み合わせて時系列データ全体を言葉で表現する手法を各々提案している [9]．

これらの手法を用いることで，時系列データと記事のアラインメント行うことも考えられるが，日付や変動の曖昧な表現についてはほとんど考慮されていない．

本研究は，自然言語表現の中でも期間と変動の曖昧な表現に着目し，グラフの各始点と終点との一致度を算出することでアラインメントを試みる点が先行研究と異なる．

時系列データと文書を組み合わせるには，文書中に含まれる時系列データに言及した文章から，該当する時系列データの箇所を特定する必要がある．例えば「中国経済の先行き不透明感から日経平均株価は1月7日に18000円を割り込んだ」という文章の場合，日付に関する表現（i.e.,「1月7日」）と値に関する表現（i.e.,「18000円」）から対応する時系列データ（i.e., 日経平均株価データ）の該当する箇所を特定し，時系列データと文書の紐付け処理（以下，アラインメントと記す）を行う．しかし，文書中の表現は必ずしも明確な数値で記述されるわけではなく,「中国経済の先行き不透明感から，日経平均株価は 1月初頭 に 大きく下落 した」のように，下線部のような曖昧な表現を用いて期間や値の変動が記述される場合も多い．このような文章からアラインメントを行うには，これらの表現を解釈し，該当する箇所を特定する必要がある．本稿では，このような文章に含まれる曖昧表現として日付に関する曖昧表現（e.g.,「中旬」「初頭」）と値の変動に関する曖昧表現 (e.g.,「上昇」「下落」）のふたつに着目し，これらから時系列データの該当箇所を特定する方式について検討する．

4.2 提案手法

図4に，グラフで表現した時系列データとそれに言及したふたつの文章を示す．図4中の (α) と (β) の文書では変動に関してどちらも「下落した」と言う表現が用いられているが，(α) の文章が紐付けられるべきは図4-Aの領域であり，(β) の文章が紐付けられるべきは図4-Bである．このようなアラインメントを行う場合，文書に含まれる変動に関する表現と日付に関する表現の両方を満たす箇所を時系列データから特定することになる．ここで,「下落した」という表現を解釈する場合，(β) のような場合には時系列データの値は必ずしも単調減少するわけではなく，その途中で一旦上

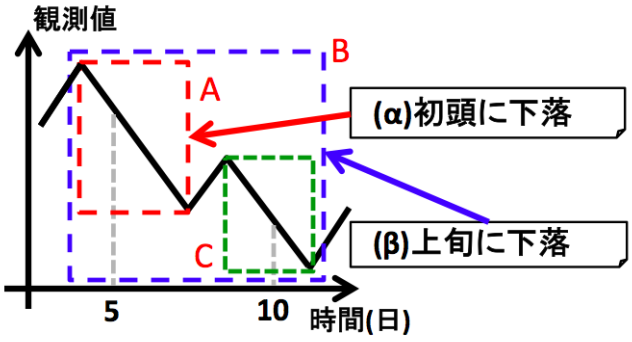

図 4: 時系列データと文書の対応

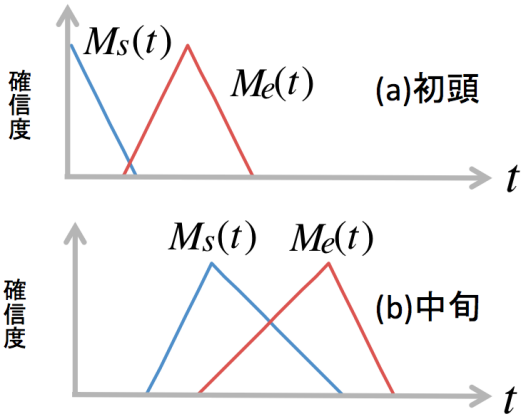

図 5: 期間のファジィ集合

昇に転じている場合も該当するため，単調減少している箇所のみをアラインメントの候補にするのではなく，(1) 領域内で単調減少している割合が高い，(2) 領域内で始点の値が最も高く終点の値が最も低い，というふたつの条件を満たす領域を選択する必要がある．本研究では，時系列データを極値で分割し，任意の二つの極値の間の時系列データを候補としてアラインメント対象の文章に照らして「変動に関する一致度」と「日付に関する一致度」のふたつを算出し，それらの値の積が最も高くなる期間をアラインメントを行う最尤期間とする．

この方式では，任意のふたつの日付 $d_A, d_B (d_A < d_B)$ で挟まれた期間 $[d_A, d_B]$ の時系列データを対象として，変動に関する一致度 C_f と期間に関する一致度 C_d を求め，それらの積を文書との一致度 C とする．C_f は $[d_A, d_B]$ で変動に対する曖昧表現を満たす区間の割合である．日付 d の時系列データの値を $f(d)$ とすると，例えば「下落した」の場合は，$[d_A, d_B]$ において $f(d_A)$ が最大かつ $f(d_B)$ が最小という条件の下で，$f(d) < f(d+1)$(但し $d_A \leq d < d_B$) となる部分区間の割合を C_f とする．また，期間に関する一致度 C_d は，日付に関する曖昧表現を日付の始点と終点を表すファジィ集合を各々 $M_s(d)$，$M_e(d)$ として（図 5 参照），$C_d = M_s(d_A) \times M_e(d_B)$ で算出する．これらから，$C = C_f \times C_d$ を算出し，それが最も高い値となる $[d_A, d_B]$ を最尤区間とする．

4.3 評価と考察

提案手法を用いたシミュレーションとして，時系列データとしてマネースクウェア・ジャパンの HP(http://www.m2j.co.jp/market/historical.php) で公開されている米ドル/円の為替レートのうち 2014 年 10 月 1 日から 11 月 28 日までのデータを，それに言及する文章として WEB 上から取得した（A）「11 月上旬は上昇した」(B)「10 月末は下落した」の 2 つの文章を用いて，これらのアラインメントを試みた．文章（A）に対するアラインメント候補上位 5 件の期間及び一致度 C を表 1 に，各々の文章の最尤区間をグラフで示したものを図 6 に示す．図 6 の結果から目視ではあるが，それぞれの文章が各時系列データの該当する箇所に対応付けられていることがわかる．しかし，一致度の算出するにあたり，「末」や「初頭」など短い期間に関する一致度が全体的に低くなるとともに，候補数が極端に少なる傾向が見受けられた．これは，ファジィ集合に設定する日付を変更すると確信度が大きく変化するからである．今後は，より多様な時系列データと文章を対象としてアラインメントを行い，提案手法の精緻化を目指す．加えて，システムによって提示されたアラインメント候補の妥当性を検証するために被験者実験を行う必要があると考えている．これにより，人間が言語を介して行っている知的情報処理により近いアラインメント候補を求めることができる．

表 1: 文章（A）に対するアラインメント候補

順位	期間	一致度
1 位	11/1 - 11/11	0.80
2 位	11/2 - 11/11	0.70
3 位	11/4 - 11/11	0.60
4 位	11/1 - 11/14	0.54
5 位	11/2 - 11/18	0.47

5 おわりに

本稿では，時系列データの探索的分析を支援するシステムの実現に向けて時系列データと新聞記事をアラインメントする方式を提案した．この方式をシステムに組み込むことにより，効率的に時系列データと文章を対応付けることができる．また，これまでの取り組

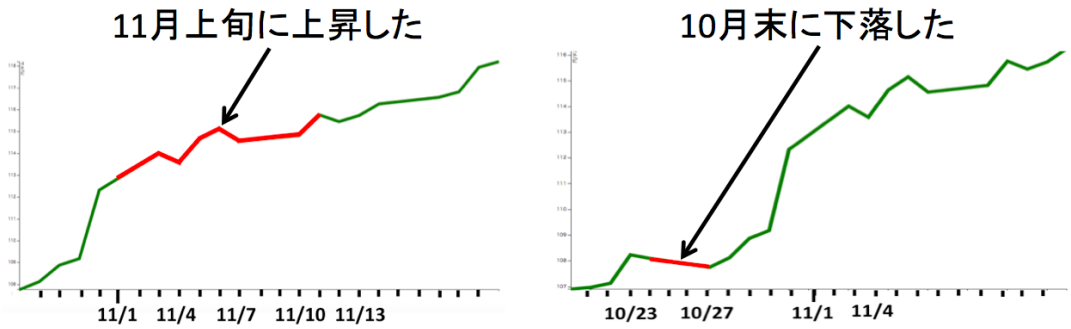

図 6: 推定された最尤アラインメント期間

みとこれから目指すシステムの全体像について述べた．さらに，インタフェースに地図と記事に相互にアクセスできる機能を実装した．これにより，興味を持った国に関する統計量の変化の理由や背景を把握したり，記事の出来事の影響や規模を把握したりできる．現在のプロトタイプシステムは,「エボラ出血熱」というトピックのみしか対応していないが，同じ要素を持つ「熊本地震」や「台風の被害」などのトピックにも応用することができるため，他のトピックも分析できるようにするつもりである．今後の展望としては，システムに必要な新聞記事のデータを Web から自動で収集する仕組みを検討している．具体的には，クローラとスクレイピング技術を用いてニュースサイトから記事の本文を抽出することを考えている．これにより，熊本地震のようなトピックを分析する人がリアルタイムな情報を見て分析を行うことができたり，システムの自動化に繋がったりする．さらに，グラフペインに複数の統計グラフを描画する機能を実装しようと考えている．時系列データの分析をする場面では，1 つのデータを見るだけでなく複数のデータを比較し，相関や違いを見ることが重要である．この機能により，ユーザは複数の統計データを比較することが可能になる．

謝辞

本研究の遂行にあたり，文部科学省科学研究費 (課題番号：15H02780) の助成を受けた．記して謝意を表す．

参考文献

[1] 藤本和則, 木村 陽一, 松下 光範, 庄司 裕子: 意思決定支援とネットビジネス, オーム社 (2005)

[2] Naito, S., Matsushita, M.: Supporting Consecutive Data Exploration by Visualizing Spatio-temporal Trend Information, *in Proceedings of the 2015 Conference on Technologies and Applications of Artificial Intelligence*, pp. 227–231 (2015)

[3] 内藤峻, 松下光範: 時空間動向情報を対象とした探索的データ分析のための可視化インタフェースの提案, ARG 第 6 回 Web インテリジェンスとインタラクション研究会, No.6, pp. 31–36 (2015)

[4] 松下光範，加藤恒昭：言語情報と数値情報の相補的利用を目指した可視化手法，第 21 回人工知能学会全国大会，3H8-3 (2007)

[5] 小林一郎: グラフ情報の自然言語処理に関する研究 日本ファジィ学会誌, Vol.12, No.3, pp.406–416 (2000)

[6] 小林 一郎, 渡邉 千明, 奥村 奈穂子: グラフとテキストの協調による知的な情報提示手法: 日経平均株価テキストとグラフの提示を例にして, 情報処理学会論文誌, Vol. 48, No. 3, pp.1058–1070 (2007)

[7] Saif Ahmad, Paulo C F de Oliveira, Khurshid Ahmad: Summarization of Multimodal Information, *Proc. 4th International conference on Language Resources and Evaluation*, pp.1049–1052 (2004)

[8] Sarah Boyd: TREND: A System for Generating Intelligent Descriptions of Time-Series Data, *Proc. IEEE International Conference on Intelligent Processing Systems* (1998)

[9] 馬野 元秀, 小泉 尚之, 篠原 貴之, 瀬田 和久: 全体的傾向と局所的特徴に基づく時系列データの言葉による表現, 第 22 回ファジィ システム シンポジウム 講演論文集, pp.343–346 (2006)