

係り受け解析を用いたブログユーザの行動・興味に関する時系列推移 3次元可視化システム*

伊藤 正彦^{†a)} 吉永 直樹^{†b)} 豊田 正史^{†c)} 喜連川 優^{†d)}

3D Visualization System for Temporal Changes in Blog Users' Activities
and Interests Based on Dependency Analysis*

Masahiko ITOH^{†a)}, Naoki YOSHINAGA^{†b)}, Masashi TOYODA^{†c)},
and Masaru KITSUREGAWA^{†d)}

あらまし ブログは、ユーザが個人の行動や興味、主観的意見などを時間情報とともに記録する媒体として近年広く用いられ、マーケティングや社会分析の観点から重要なデータとなってきた。本論文ではブログに記述された対象に対する行動及び興味の時系列推移を3次元可視化・探索する話題探索システムを提案する。提案システムでは、まず、ブログアーカイブから対象に関する行動・興味の記述（例：チョコレートを食べる）に係り受け解析を用いて抽出し、月単位の粒度でデータベース化する。更に、入力キーワードに関する行動・興味の記述とその出現頻度の時間変化を可視化し、そのトレンド及び詳細を対話的に探索可能にする3次元システムを提案する。実際に、過去5年分からのブログアーカイブを用いて提案システムを構築し、いくつかの応用事例を用いて提案システムの有用性を示す。

キーワード ウェブ、自然言語処理、係り受け解析、3次元情報可視化、時系列情報可視化

1. ま え が き

ブログやマイクロブログなどの Consumer Generated Media (CGM) の普及に伴い、ブログユーザは自身の興味、行動、主観的意見を即座にかつ簡単にウェブ上に記述することが可能になってきている。これら、時間・社会状況とともに変化するブログユーザの生の声は、製品、人物、政策等の評判調査など、マーケティング、社会分析の観点から重要なデータ [1] である。ウェブ上のデータは多様でかつ膨大なものであるため、多視点から長期にわたる人々の行動・興味の移り変わりを分析し、可視化によりその結果を俯瞰可能にするシステムへの要望が強まっている。

このような背景から、kizasi.jp [2], BlogScope [3],

BlogPulse [4] 等の CGM 分析サービスや blog-Watcher [5] のようなシステムが提案されている。これらのサービスでは、注目キーワードを抽出し、その出現頻度の変化をバーチャート等で可視化し、ユーザに提示する。これらの機能により、ユーザは注目する対象がいつごろ話題になったかを知ることができるが、どのような文脈で話題になっているかを把握することは困難である。kizasi.jp 及び BlogScope のように選択されたキーワード（例：「鳩山首相」）に対して共起語（例：「沖縄」、「オバマ」、「ルーピー」など）を関連語として提示するサービスもあるが、それぞれがどのような文脈で現れたか分からない上、無関係の語が含まれることがあり、実際の文脈を類推することは困難である。関連語が現れた文脈を理解するためには、多くの場合個々の CGM の投稿までたどって実際の記述内容を読む必要が生じる。

これに対し、我々が提案するシステムでは、ブログアーカイブに対して係り受け解析を行い、注目する単語と関連語が直接共起する文脈をイベントとして収集

[†] 東京大学生産技術研究所, 東京都

Institute of Industrial Science, The University of Tokyo,
Tokyo, 153-8505 Japan

a) E-mail: imash@tkl.iis.u-tokyo.ac.jp

b) E-mail: ynaga@tkl.iis.u-tokyo.ac.jp

c) E-mail: toyoda@tkl.iis.u-tokyo.ac.jp

d) E-mail: kitsure@tkl.iis.u-tokyo.ac.jp

* 本論文はデータ工学研究専門委員会推薦論文である。

(注1): 本論文では、トリーをツリーと呼ぶ。

し、まとめて可視化することでユーザに提示する。具体的には、動詞とそれに係る名詞からなる、「対象に関する行動の記述（例：チョコレートを食べる）」若しくは「ブログユーザが興味をもっている状態・状況の記述（例：新型インフルエンザが流行る）」をイベントとして抽出する。単純な共起単語ではなくイベントという関係を通して抽出された関連語のみを用いるため、表示されるノイズ成分は軽減される。更に、イベントの情報を段階的に詳細化しながら探索する仕組みにより、興味をもったイベントに関して、誰が、どこで、何を、なぜ、などの行動・興味の詳細を調査することが可能になる。これらにより、そのキーワードがどのような文脈で話題になっているかに関して、より容易に理解することが可能となる。

与えられたキーワードに関する人々の行動・興味の移り変わりを文脈も含めて理解するためには、単に関連単語群とその頻度変化を可視化するだけでは不十分である。そのためには、文脈情報を含むイベント群の全体像を俯瞰し、それらの時間変化を可視化する環境が必要になる。本研究では、与えられたキーワードに関係するイベントとその詳細情報を集約し、それらの構造をツリー表現により可視化し俯瞰可能にする。また、3次元空間の一軸を時間軸として用いることで、イベント群の構造の時間変化と各イベントの出現頻度の時間変化を俯瞰可能にする。このように、係り受け関係の集合を可視化により俯瞰可能とした研究は、我々の知る限り他に存在しない。

これらを実現するにあたり、提案する可視化システムでは以下の三つの基本機能を実現する。

(1) 入力キーワードに関する特定年月のイベント群とその文章構造を、キーワードを中心としたツリー表現を用いて可視化する。各イベントを入力キーワードに付属する格助詞で分類することで、係り受けの意味的關係（行動の動作主、対象、場所など）によりイベントを分類表示する。また、それらのイベントの詳細情報をサブツリーとして可視化し探索可能にする。

(2) ユーザが様々な観点から情報を比較・観測できるようにするため、異なる年月に関するイベント群を別々のツリーとして可視化する。また、異なるキーワードに関するイベント群も別々のツリーとして可視化する。

(3) 3次元空間の1軸を時間軸として用いることにより、時間軸に沿った出現イベント群の変遷を観測可能にする。また、複数ツリーを時間軸上に並べるこ

とで出現イベントの違いを俯瞰可能にする。更に、各イベントに関する出現頻度を時間軸方向に串刺しに可視化し、その頻度変化を俯瞰可能にする。

これらの基本機能により、ブログに記述された対象に対する行動・興味の時系列推移を可視化・探索する話題探索システムを実現する。本研究では、実際に過去5年分からなるブログアーカイブを用いて、いくつかの応用事例を示すことで提案システムの有用性を示す。

本論文の構成は以下のようになっている。まず、2. でイベントの定義とその抽出手法を述べるとともに、イベントデータベースについて説明する。次に、3. で3次元可視化システムについて述べ、4. では、2. と3. で提案した仕組みを用いた応用例を紹介する。5. では、共起語、係り受け関係などの単語間関係を用いた分析システム及び時系列変化の可視化システムに関して従来手法の具体例をいくつか紹介しその問題点を明らかにする。最後に、6. で全体のまとめを行う。

2. ブログアーカイブに対する係り受け解析とイベントデータベース

筆者らの研究室では、過去5年間、3億5000万エントリに及ぶ日本語ブログアーカイブの構築を進めてきている。膨大なブログエントリ集合から入力されたキーワードに対するイベントを高速に検索可能にするため、蓄積された全ブログエントリに対してあらかじめ係り受け解析を行い、ブログユーザの行動・興味に関する記述をイベントとして収集し月単位でイベントデータベースを構築した（図1）。

2.1 ブログアーカイブからのイベント抽出

イベントデータベースの構築にあたり、まず図1(a)に示すように月ごとのブログエントリの各文章に対して係り受け解析を行い、依存構造木を構築し、それらからイベントを抽出する。

提案システムでは、日本語係り受け解析システム J.DepP^(注1)により係り受け解析を行う。J.DepPには、複数の統計的係り受け解析アルゴリズムが実装されているが、本研究は、線形時間アルゴリズム[6]を利用し、更にアルゴリズム内で係り受けの有無の分類に用いる分類器を素性列トライ[7]を用いて高速化することで、大規模ブログデータの解析を可能とした。

J.DepP が出力した係り受け解析結果のうち、動詞

(注1) : <http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/>

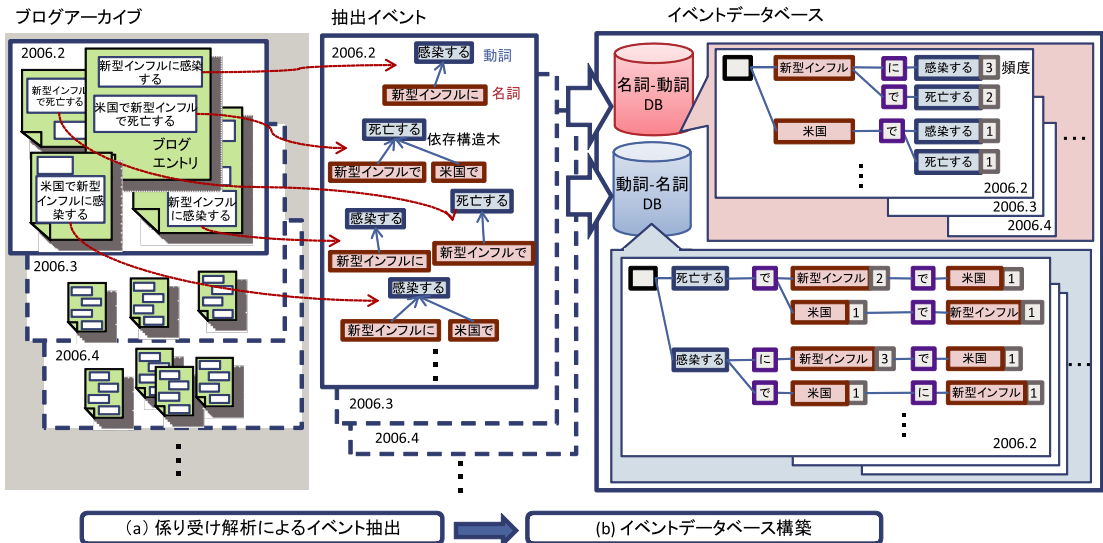


図1 ブログアーカイブ中のブログエントリーに対する係り受け解析を用いたイベントデータベースの構築

Fig. 1 The overview of constructing event databases from blog posts in a blog archive by using dependency analysis.

と動詞に直接係る名詞の間の係り受け関係を一つのイベント（対象＋行動）として抽出する．イベント中の各動詞は，ブログユーザの行動や彼らが興味をもっているであろう状態・状況を表す．更に，名詞はその動詞を修飾することで行動や興味の特徴をより詳細に表現する．ユーザは，特定のイベントの詳細をイベントに含まれる動詞とそれに係る名詞の集合から要約し理解することが可能になる．

例えば，「新たに13歳の少女が新型インフルエンザに感染した」という文に対する係り受け解析結果は図2の依存構造木で表現されるが，このうち太線で囲まれた関係のみをイベントとして抽出する．図3では「米国で児童が新型インフルで死亡したので，不安が広がっている」という，複文に対する依存構造木を示す．動詞と動詞に直接係る名詞の間の係り受け関係を一つのイベントとして抽出するため，赤の太線で囲まれた関係と青の破線で囲まれた関係は，独立した二つのイベントとして抽出されることに注意されたい．また，図3の例は，係り先に曖昧性を含む文になっている（「米国で」は統語的には「死亡したので」若しくは「広がっている」に係り得る）．J.DepPは，統計的構文解析器であり，このような曖昧性を含む文に対しても，[6]で示すアルゴリズムに基づき，一意に係り受け

関係を決定し，依存構造木を出力する^(注2)．このようにして文から抽出されたイベントを月ごとに集約し，その頻度情報（その月でそのイベントを抽出した文の総数）とともに2.2で述べるイベントデータベースに格納する（図1(b)）．

本研究では文節単位の係り受け解析器を用いており，否定や継続など助動詞で表現される動詞のモダリティは，図2や図3の依存構造木のように，動詞と同一の文節に含まれる．そのため，係り受けにおいて係り先にモダリティが含まれる場合（例：新型インフルエンザに感染したくなかった）は，動詞のモダリティを無視して動詞のみの係り受け（新型インフルエンザに→感染する）を考えるか，モダリティも含めた係り受け（新型インフルエンザに→感染したくなかった）を考

（注2）：ただし，依存構造木内の各係り受けについては，誤りの係り受けを含む場合があり得る．評価用コーパス[8]における評価では，係り受け単位で精度91.80%と同データでの最高精度91.96%[9]に匹敵する精度となっている．解析誤りの原因としては，連用修飾・複文・並列句など複雑な構文構造に起因する統語的な曖昧性[10]が多数を占めており，提案システムが扱う格助詞を介した係り受けの正誤への影響は限定的である．仮に，解析結果に誤りの係り受けがわずかに含まれたとしても，集約した解析結果における高頻度のイベントの順位への影響は無視できると期待できるため，イベントの総体的な振る舞いを俯瞰するという提案システムの用途に対して，用いた構文解析器の解析精度は十分と考える．

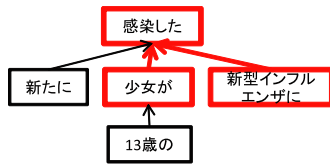


図 2 「新たに 13 歳の少女が新型インフルエンザに感染した」という文に対する依存構造木

Fig. 2 A dependency structure for a Japanese sentence: "Aratani 13-sai-no shoujo-ga shingata-influenza-ni kansensita (a 13-year-old girl newly caught a swine flu)". Dependency relations with thick line width form an event.

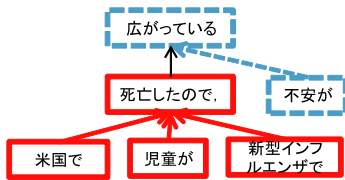


図 3 「米国で児童が新型インフルで死亡したので、不安が広がっている」という文に対する依存構造木

Fig. 3 A dependency structure for a Japanese sentence: "Beikoku-de jidou-ga shingata-influenza-de shibou-sita-no-de, fuan-ga hirogatte-iru (A child dead of swine flu in USA, so uneasiness spread.)".

えるかは、(実装上) 選択の余地がある^(注3)。提案システムでは、動詞のモダリティを全て無視したイベントデータベース、及び特に区別が重要と考えられる「否定」「受身」「使役」「願望」に関するモダリティを考慮したイベントデータベースを構築している。なお、本論文では以降、動詞のモダリティを無視したイベントデータベースを用いて各事例を解説する。

各イベントにおいて、動詞に対する名詞の意味的関係（行為の動作主や、対象など）を名詞に付属する格助詞で分類し、併せてデータベースに収納する。分類に用いた格助詞としては、高頻度で、かつ動作主、経験者、道具、対象、場所、時間など、イベントに関する重要な情報を示すのに主として用いられる{を, と, に, が, で}のみを考慮した。このようにイベントを分類することで、ユーザは検索結果中の膨大なイベント群から興味の対象となる意味的関係をもつイベントを、比較的容易に探索可能になる。

2.2 イベントデータベース

今回、ブロッガーカブから約 150 万種類の名詞句、約 11 万種類の動詞句から構成される延べ約 12 億のイベントを獲得した。本システムでは、抽出されたイベントに対して、2 種類の係り受け解析結果の

データベース（イベントデータベース）を構築した（図 1 (b)）。データ構造としては簡潔データ構造に基づくトライ[11]を用いた。

名詞-動詞データベース：注目したいキーワードをクエリとして、月ごとのイベント及びその頻度を検索可能にする。この場合、名詞（キーワード）から、その名詞の係り先となり得る動詞をイベントとして検索し、検索されたイベントを構成する係り受けを含む文の総数を頻度情報として返す。図 1 の例では、キーワード「新型インフルエンザ」をクエリとして与えた場合、データベースは「に 感染する」とその頻度 3、及び「で 死亡する」とその頻度 2 を結果として返す。データベース中のトライ木では各イベントは名詞、格助詞、動詞の順序で連結された文字列として格納され、与えられた名詞を根とするサブツリーを返す。

動詞-名詞データベース：動詞若しくはイベントをクエリとして、月ごとのイベント及びその頻度を検索可能にする。動詞をクエリとする場合、データベースは与えられた動詞に係る名詞を検索結果として返す。一方、イベントをクエリとして与えた場合、データベースは与えられたイベント中の動詞に係る名詞を検索結果として返す。図 1 の例では、イベント「新型インフルエンザ で 死亡する」をクエリとして与えた場合、検索結果は「米国 で」となる。データベース中のトライ木では動詞とその動詞に係る（格助詞、名詞）が順に格納され、与えられたイベント（動詞、（格助詞、名詞））による経路以下の部分木を格助詞及び名詞各 1 語分の深さで返す。

このように、各イベントを文字列としてトライに格納することで、キーの冗長性（図 1 右）を大きく削減できることに注意されたい。更に、各単語をデータ

(注3)：まず、モダリティを全て無視した場合には、「新型インフルエンザに感染しなかった」や「新型インフルエンザに感染した」、また「新型インフルエンザに感染したくない」などは、全て「新型インフルエンザに」→「感染する」という構造をもつ依存構造木となり、同じイベントとして扱われる。モダリティを無視することにより、同じ動詞と名詞に関連するイベントを集約し頻度を算出することが可能になる。これにより、新型インフルエンザへの感染に関する話題そのものに焦点を当ててイベントの隆盛を把握することが可能になる。一方、モダリティを考慮することにより、「新型インフルエンザに感染しなかった」や「新型インフルエンザに感染した」、または「新型インフルエンザに感染したくない」などは、「新型インフルエンザに」→「感染しなかった」、「新型インフルエンザに」→「感染した」、または「新型インフルエンザに」→「感染したくない」のように、動詞のノードが異なる依存構造木となり、別々のイベントとして扱われる。これにより、話題に関する詳細な可視化が可能になる一方、細かい粒度で動詞を区別する副作用として、各イベントの頻度が分散し、流行している話題が見つけにくくなるという問題が発生する。

ベースでの出現回数を考慮して可変長バイト符号化により圧縮することで、データベースのサイズを削減している。これらの工夫により、上記のイベントデータベースはオンメモリで利用できるようになっている。

3. 3次元可視化システム

2. で述べた係り受け解析結果データベースに対して、入力キーワードとそれに関連するイベント群の月ごとの変化及び各イベントの頻度変化を探索可能にする3次元可視化システムを構築する。提案する3次元可視化システムでは以下の機能を実現する：(i) 入力キーワードに関する特定年月のイベント群の可視化、及び各イベントの詳細情報可視化、(ii) 異なるキーワード若しくは異なる年月から構成される複数イベント群の比較、(iii) 各イベントの時系列頻度変化の可視化。

可視化部品として TimeSlice [12] 及び TimeFlux [13] を用いることで上記の機能を実現した。TimeSlice は3次元空間の時間軸に沿って配置可能な時系列グラフ表示部品である。TimeFlux は TimeSlice 上の各ノードに対し月ごとの頻度変化を球の列として可視化する部品である。インタラクティブな3次元アプリケーション構築環境として IntelligentBox [14] を用いている。

図4の例は、複数の携帯キャリアに関するイベント群の時系列変化を可視化、比較したものである。TimeSlice を用いることで、イベント（対象と行動・状態の関係）をツリー表示でき（図4(a)）、更に、ノードを選択し展開することで、誰が、どこで、何を、などイベントの詳細情報が探索可能になる（図4(b)）。TimeSlice は3次元空間中の時間軸に沿って配置され、時間軸の位置に対応したイベント群の表示を可能にする。ユーザは TimeSlice をドラッグすることで表示時間をシームレスに変更させ、イベントツリーの構造変化をアニメーション表示させることが可能となる。TimeSlice は時間軸上に複数枚並べて追加することが可能であり、これにより異なる月ごとのイベントツリーの違いを観測できる（図4(c)）。更に、上下に TimeSlice を並べて追加することで異なるキーワードから検索されたイベント群を比較することが容易になる（図4(d)）。TimeFlux を用いることで、各イベントに対する月ごとの頻度を時間軸方向に連結して一覧表示できる。これにより、イベントの頻度変化を直感的に観測可能になる（図4(e)）。

以下で、各機能の詳細を解説する。

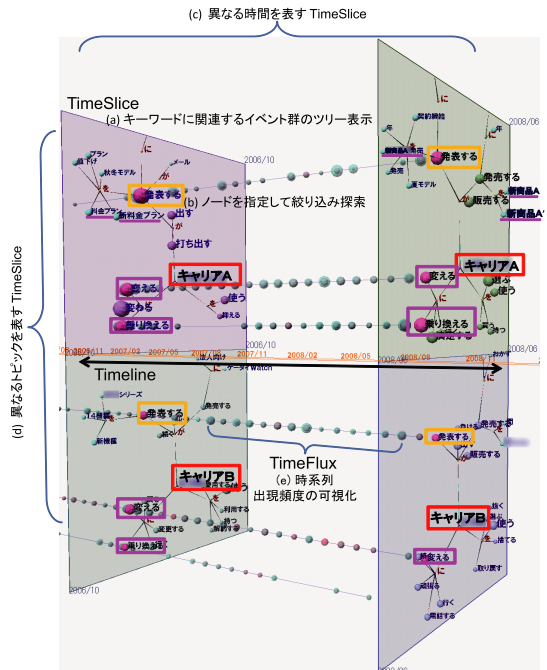


図4 TimeSliceによる複数キーワード、複数時間からなるイベント群の可視化、詳細情報探索、及び TimeFlux によるイベント頻度変化可視化

Fig. 4 Multiple TimeSlices for comparing events in different months and for comparing events on different topics.

3.1 TimeSlice によるイベント可視化

名詞-動詞データベースは、図1右上で示すようにクエリ（名詞）に対して月ごとにその名詞を含むイベント集合を木構造として返す。TimeSlice は、この検索結果を入力とし、時間軸上で指定された月に関するイベント群を、入力キーワード（名詞）を中心とし、その下に格助詞、各格助詞の下に動詞が配置されたツリー表現で可視化する。

図5(a)の例では、入力キーワード「新型インフルエンザ」を中心として、その周りに格助詞、動詞がツリー表示されている。中心となるキーワードからツリーのパスをたどることで、例えば「新型インフルエンザに感染する」といったイベントの内容を解釈可能である。見たい格助詞のみを表示するフィルタリング機能も用意している（図5では{で、に、が}のみ表示）。

イベントは各月ごとに集約されており、各格助詞の下には指定された月の出現頻度上位 n 個（ n は任意）のイベントのみを表示する（図5では上位五つ

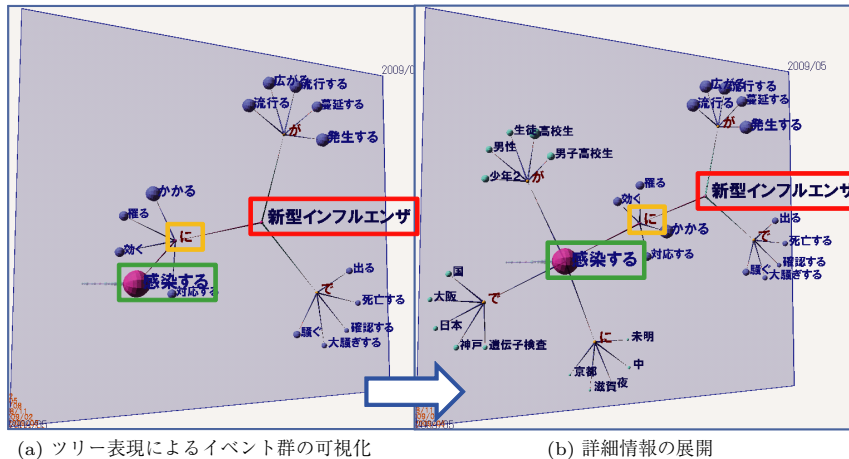


図 5 ツリー表現によるイベント群の可視化とノードの選択・展開による付帯情報可視化
Fig.5 Visualizing set of events by tree on a TimeSlice, and detailed information about a selected event.

を表示). 名詞-助詞-動詞で表現するイベントの出現頻度は、動詞に対応するノードのサイズで表現する. ノードサイズが巨大になるのを防ぐため、サイズは $r\sqrt[3]{c_i(t)}$ で計算している (r : 定数, $c_i(t)$: イベント i の月 t における出現頻度). ユーザは TimeSlice を時間軸に沿って移動させることにより、指定された年月における出現頻度上位 n 個のイベント及びそれぞれの頻度の時間変化をアニメーション表示させることができる. ノードの色は要素ごとに任意の色を設定できる.

3.1.1 イベント詳細探索インタフェース

可視化結果から新たな検索要求を作成し、段階的に情報探索を行うインタフェースについて説明する. ノードを選択することで詳細化するイベントを限定し、それに対する付帯情報を検索し展開表示を行う. これにより、選択したイベントに関して更に詳細な「誰が、どこで、何を」などの情報を一覧できる. 図 5 の例では、「感染する」というノードを選択することで、「新型インフルエンザに感染する」というイベントを選択し、動詞-名詞データベースから付帯情報を検索・可視化している. この例では「高校生が」「神戸で」などを選択イベントの経路以下の部分木として動詞-名詞データベースから取得し、得られた付帯状況を「感染する」の子ノードとして可視化している.

3.2 複数 TimeSlice による比較

3.2.1 複数キーワードの比較

複数のキーワードに対する検索結果を上下の TimeSlice に分けて表示する Split view [12] を用いること

で、複数キーワードに対する比較を行うことが可能になる (図 4 及び図 8). 両者の比較を容易にするために、上下の TimeSlice に対する時間軸上の位置指定の操作は同期がとられ、両者の時間は自動的に一致させることができる.

異なるキーワードにおけるイベント群の類似・差異を理解しやすくするため、両者の間での同等のイベント (中心となる名詞、同じ格助詞、及び同じラベルのついた動詞を表すノードからなるパス) を各 TimeSlice 上で同じ位置に配置する.

3.2.2 複数時間の比較

各検索結果は年月情報をもっており、TimeSlice 上には指定した年月に関する結果のみが表示される. ユーザは、時間軸 (Timeline) をクリックすることで、指定した時間上に任意の数の TimeSlice を追加することが可能である. これにより、異なる複数時期による人々の動向及び興味の違いの詳細を比較できる (図 4 及び図 9). なお、複数キーワードに対する TimeSlice を上下に表示している場合、追加操作は上下同時に行われる.

複数時間におけるイベント群の類似・差異を理解しやすくするため、異なる時間を表す隣り合う TimeSlice 間では同じイベントは同じ位置に表示される.

3.2.3 複数 TimeSlice に対するタイル表示

3 次元空間の一軸を時間軸として扱う利点としては、構造変化の時間感覚が分かりやすく、変化の全体像の俯瞰がしやすい点が挙げられる [15]. TimeFlux を併

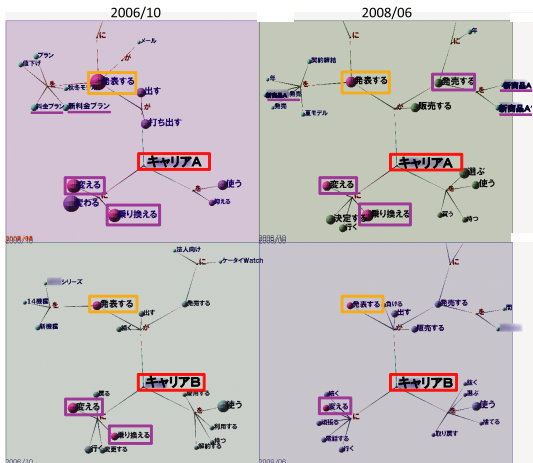


図 6 タイル表示を用いた比較インタフェース

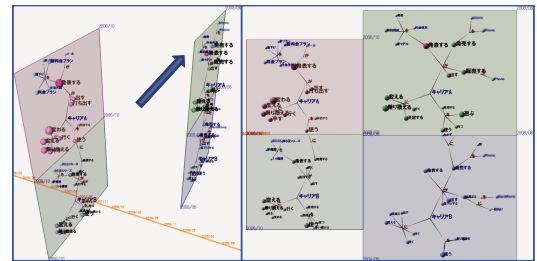
Fig. 6 Tiled view.

用することにより、構造変化と併せて個々のイベントに関する頻度変化の俯瞰も可能になる。

一方、3次元空間に複数の2次元パネルを重ねて並べる手法(2.5次元可視化)では、以下の問題が生じることが指摘されている。まず、パネルが重なりあう[16]、また、斜め横から見る場合、遠近法のため奥行方向の面積が狭くなり見づらくなる[15]、[16]、更に、遠近法の影響で時間軸上のTimeSliceの位置によって画面上を占める面積や表示面(表裏)に違いが生じる[15]。

本論文では、複数時間を表すTimeSliceに対するこれらの問題に対応し、かつ、TimeSlice間の差異を詳細に観測可能にするため、タイル表示機能[12]を導入した(図6)。タイル表示は3次元空間で実現され、ユーザはシームレスに通常の3次元表示とタイル表示を切り換えることが可能である。表示方式を自由に切り換えることで、ユーザは3次元表示で全体を俯瞰し、アニメーションなどにより特徴的な変化を見つけ、興味のある領域を表示した状態でタイル表示に切り換え、その領域に関する詳細を比較し探索する、といった操作が可能になる。図6は図4をタイル表示に切り換えた例である。

表示方法の切替は、TimeSliceを奥行方向にスライドさせた状態(図7(a))で、ユーザ視点を時間軸方向からTimeSliceを見るように移動し(図7(b))、正射影モードで描画することで実現する(図6)。通常の透視投影では、図7(b)のように手前のTimeSliceと奥にあるTimeSlice上のツリーが異なるサイズで描画



(a)TimeSliceの奥行き方向へのスライド (b)視点を時間軸方向へ移動

図 7 3次元空間におけるタイル表示へのシームレスな切替

Fig. 7 Changing a normal 3D view to a tiled view in a 3D space.

される。正射影で描画することで、異なる時間を表すTimeSlice上の同じノード、若しくは異なるキーワードを表すTimeSlice上の同等のノードの位置は完全に一致する。同様の仕組みを用い、奥行方向のスライドをしないことにより、複数時間の状態を重畳表示して比較することも可能である。

3.2.4 ノードレイアウト

複数のTimeSliceは異なるキーワード及び異なる時間に属するイベント群を同時に可視化する。これらのTimeSlice間では同じ若しくは同等のイベントは同じ位置に表示される。本研究ではDiehlら[17]の手法を応用し、全TimeSlice上のその時点で表示される全ノード及びエッジを集約したツリーを生成し、その集約ツリーに対するレイアウト計算を行う。各TimeSliceは、レイアウト計算された集約ツリーから、指定された時間及びキーワードに関するノードとエッジのみを選択し表示する。3.2.1で述べたような、異なるキーワードから構成されたイベント群間における同等のイベントを除き、異なるキーワードに関するツリーは、それぞれ独立にレイアウト計算される。

TimeSliceに表示されているツリーは力学的アプローチ[18]に基づいたレイアウトを行っている。力学的アプローチは、グラフにおけるノードを質点、エッジをばねとする力学モデルのシミュレーションを用いてグラフをレイアウトする手法である。エッジで結ばれた2点間には引力が働き、全ての2点の間には斥力が働く。

3.2.5 複数TimeSliceに対するインタラクション

ユーザは対話的にノードをドラッグしノードの位置を変更することができ、ノード間の関係を確認できる。更にTimeSlice上の表示領域に対するズームイン

及びパンニングの操作も対話的に行える。これらの対話的な操作を行う際にも、異なる TimeSlice 間の同じ若しくは同等なノードの位置は自動的に同期する。更に、表示領域に対するズーム、パンの操作も複数 TimeSlice 間で同期する。

TimeSlice 上ではノードを選択することにより、選択されたイベントはハイライト表示される。この際、一つの TimeSlice で選択されたイベントに関して、他の TimeSlice 上の同じ若しくは同等のイベントは自動的にハイライト表示される。

同様に、一つの TimeSlice で実行された詳細情報表示のための展開操作も、自動的に他の TimeSlice に派生する。これにより、複数の TimeSlice に同じ動詞表現からなるノードがある場合、同じ条件で比較を行うことを可能にしている。図 4 の場合、一つの TimeSlice に対して展開操作（例えばイベント「キャリア A が発表する」を展開）を行うことで、異なる時間を表す TimeSlice でも同じイベント（「キャリア A が発表する」）に対して展開操作が自動的に行われる。同様に、異なるキーワードを表す TimeSlice 間では同等のイベント（例えば、「キャリア B が発表する」）に対する展開操作が自動的に行われる。

3.3 TimeFlux による時系列頻度変化の可視化

選択したイベントに関して月ごとのイベント出現頻度を時間方向に連結した球の列として可視化することで、どの時期にそのイベントが話題になったか、他のイベントとの傾向の違い、流行の周期性等が直感的に理解可能になる。この球の列を TimeFlux [13] と呼ぶ（図 4(e)）。TimeFlux は、ユーザが選択したイベント（ノード）をクエリとしてイベントデータベースから取得した月ごとの頻度情報を球の大きさで表現し、選択したノードの時間軸上に可視化する。ユーザは TimeSlice 上の任意のノードを選択することで複数のイベントに関する TimeFlux を表示できる。

図 8 の例では、上部では「新型インフルエンザに感染する/かかる」に関する TimeFlux を、下部では「インフルエンザに感染する/かかる」に関する TimeFlux を表示している。この図から「新型インフルエンザ」に関する TimeFlux では 2009 年 5 月ごろに頻度の急激な増加（バースト）が見られ、大流行が起こったことが分かる。また、下部の「インフルエンザ」は主に「季節性インフルエンザ」を意味しており、毎年冬期になると周期的にはやっていることが確認できる。

各ブログは解析前処理として、一人称代名詞及び終

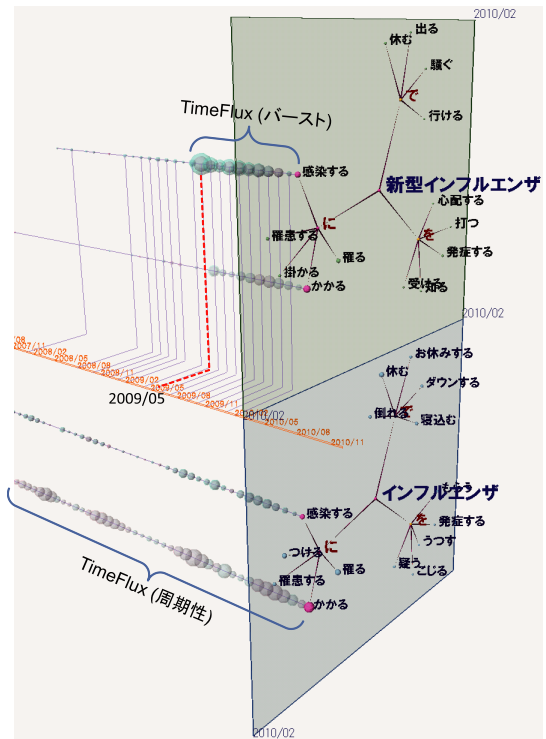


図 8 複数 TimeFlux を用いた頻度変化の比較

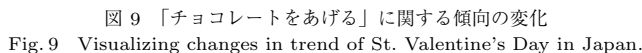
Fig. 8 Comparison of temporal changes in frequency among events about different topics by using multiple TimeFluxes.

助詞を手掛りに簡易的に男女の性別を判定 [19] しており、2. で述べたデータベースでは各イベントの頻度を男女別に集計している。本論文では男性の頻度を表す TimeFlux に青、女性の頻度を表す TimeFlux に赤を割り当てている。これにより、各イベントに対する男女の行動の違い、興味の違い、及び男女比の推移を可視化し観測、比較することが可能になる。図 9 の例では、「チョコレートあげる」というイベントに関する男女比の割合が逆転していく様子を可視化している。

4. 応用例

本章ではいくつかの応用例をその探索シナリオとともに述べ、提案システムの有用性を示す。

これらの例では、従来の共起語を可視化する話題分析システムでは読み取ることでできない、我々の提案する 3 次元可視化システムならではの可視化結果を確認できる。単純な共起単語ではなく係り受け解析結果から抽出されたイベントとその頻度を可視化することで、キーワードが使われる文脈及びその変化を容易に



TimeSlice を時間軸に沿って動かしアニメーションによって全体の変化を観察していると、キャラヤ B に比べてキャラヤ A では「乗り換える」「変える」の記述が目立つことが分かる。「乗り換える」「変える」のイベントに関する TimeFlux を表示すると、キャラヤ

これからキャリア A のマーケティング戦略は社会

1462

に大きなインパクトを与えているであろうことが見て取れる．一方、キャリア B においても何度も「発表する」の頻度が大きくなっているにもかかわらず、それが「乗り換える」や「変える」という行動にあまり大きな影響を与えていないことがうかがえる．

提案システムを用いることにより、興味のあるキーワードに対して影響若しくは重要性があると考えられる「人々の特徴的な行動（例えば「キャリア A に乗り換える」）」が多く現れた時期を発見し、そのタイミングで他にどのような現象が発生していたのかを詳細に探索することが可能になる．

4.2 バレンタインデーに関する傾向の変化可視化

図 9 は、「チョコレートを食べる」ことに関する世の中の傾向の変化を可視化した例である．はじめにキーワード「チョコレート」を入力しイベント群を可視化した．結果から「食べる」を選択し TimeFlux を表示すると、毎年バレンタインデーの時期にのみ集中してイベントが発生していることが分かる．左の TimeSlice は 2006 年 2 月の状態を表しており、右の TimeSlice は 2009 年 2 月の状態を表している．その間の TimeFlux を見てみると男女比が「#女性 > #男性」から、「#女性 ≤ #男性」へと徐々に変化してきていることが見て取れる．そこで、そのイベントに関して展開し詳細を見ると、2006 年 2 月には「女性が」「男性に」「チョコレートを食べる」例しか存在しないが、2009 年 2 月には「男子が」「女子に」「チョコレートを食べる」例が上位に出現していることが確認できる．

従来のバレンタインデーの習慣から考えると、圧倒的に「女性が」「男性に」「食べる」頻度が多くなると考えていたが、このように徐々にではあるが女性から男性にチョコレートを食べるという習慣とは逆の現象が起こっていることが観測できる．このような、各名詞の意味的役割の交替を読み取ることは、従来の共起語のみの観測では困難である．

5. 関連研究

5.1 単語間関係を対象とした検索手法

Kahoe と Gee [20] は heatmap を用いて与えられた単語に対する共起語の時間変化を可視化する手法を提案している．しかしながら、1. でも述べたように、共起語には本質的にノイズが多く含まれるため、共起語のみを見ることでイベントの分析を行うことは難しい．Resnik と Elkins [21] はユーザの入力を構文解

析し、その句構造を入力クエリとして用いた句構造検索システムを提案している．Atterer と Schütze [22] は Lucene を利用して各係り受け関係をインデックスし、係り受け検索を実現している．これらの研究では検索する係り受け関係をユーザが陽に指定する必要があるが、我々のシステムでは、係り受けのデータベースにトライ木を利用することで、インタラクティブな係り受け検索を可能としている．

Inui ら [23] は、テキストベースの経験マイニングシステムを提案している．彼らのシステムは、様々な事物（製品、公共施設、サービスなど）の利用に関する個人の経験をウェブから収集し、それらをテンス、アスペクト、極性やモダリティ情報などに基づき分類し、ユーザに提示する．我々のシステムでは、これらの利用可能な事物に関する経験を含む任意の係り受け関係をイベントとして検索可能としている点が彼らの研究と異なる．また、我々の提案する可視化手法では、特定のイベントを段階的に詳細化しながら検索することが可能となっており、ユーザが望むイベントを検索するのに必要となる労力を大幅に削減することに成功している．

5.2 構造変化の可視化

ウェブ若しくはブログのリンク構造のようなグラフ構造の時間変化を可視化する仕組みとしては、主に以下の四つの手法が提案されている．

(1) アニメーションの利用：アニメーションは時間による動的な構造変化を直感的に観察可能にし [15], [24], [25]、ユーザは変化を認知しやすくなる [15], [26], [27]．一方、状態が移り変わることで直前の状態との直接比較が困難になり、どこで変化が起こったのか、いつ変化が起こったのかを認識することが難しくなる．

(2) 3 次元空間の一軸を時間軸として利用：時間軸上に複数のグラフを並べることで複数時間の構造の違いを比較することが容易になる [15]．更に、複数グラフを俯瞰できるため変化の全体像を把握しやすくなる．一方、グラフ間の距離が離れることにより、局所的な変化の詳細比較が難しくなる場合がある．

(3) スプレッドシート風インタフェースの利用：時間を含めた様々な条件からなる複数のグラフを生成し並べて表示することにより、時間変化の全体像を把握しやすくなる [24], [28]．更に、異なる時間における構造の違いの比較が容易になる．一方、表示されているグラフ同士は実際の時間間隔にかかわらず常に等間隔

に隣接しているため、時間感覚が直感的に把握しにくいといった短所もある。

(4) 重畳表示：複数時間における異なるグラフの差異を詳細に比較することが可能になる [25], [29]。一方、この手法では、変化の全体像を把握することが難しい。

我々のシステムでは、TimeSlice [12] を用いることにより時系列グラフを可視化する上記 4 種類の手法を 3 次元空間上で統合し、シームレスに切替可能にしている。

5.3 属性値変化の可視化

5.2 で述べた (1)~(4) の手法を用いることで、ユーザはグラフ構造の変化を観測することが可能になる。一方、これらの手法ではグラフ上の個々の要素に対する属性値（例えば出現頻度）の変化を俯瞰する仕組みは用意されておらず、各要素の値がどの時期に増減したのかを直感的に把握することができない。

これらに対して、属性値の時間変化を俯瞰する手法も様々なものが提案されている。ThemeRiver [30] はキーワードごとの頻度変化を可視化可能にした。また、TIARA [31] は ThemeRiver に Tag-clouds を組み合わせることにより、トピックを構成する要素の変化をも同時に可視化可能にした。Time-tunnel [32], [33] では属性値変化のライングラフを表示すると同時に属性値同士の値の関係をレーダチャートで表示することを可能にしている。

これらの手法は頻度変化の全体像の俯瞰に優れ、各時間における属性値の違いを直感的に理解することを可能にするが、複雑な構造をもつ要素間の構造変化の可視化までは考慮していない。我々のシステムでは 3 次元空間を用いて、TimeFlux [13] と TimeSlice [12] を連携することで構造変化と頻度変化を同時に可視化することを可能にしている。

5.4 構造変化と属性値変化の同時可視化

構造変化と属性値変化の両方を同時に可視化するための解決方法の一つとしては、CMV (Coordinated Multiple Views) [34] の利用が考えられる。

月ごとのイベント群の可視化には、3.1 で紹介するようなツリー構造の可視化を用いる手法のほかに単純なテーブル表示の利用等が考えられる。例えば、著者らは各月をカラムとし、各カラムには格助詞ごとに分類したイベント上位 N 個を表示するテーブル表示インタフェースも実装済みである。各イベントに関する頻度変化や詳細情報は、バーチャートのようなイベント頻度可視化のためのウィンドウ、及び詳細情報表示用

の別ウィンドウを用意し、イベント群の可視化ウィンドウとリンクを張ることで対話的に探索可能になる。

しかしながら、CMV を用いて複数のイベントに対する頻度若しくは詳細情報を表示して情報の俯瞰やイベントの比較をしたい場合、多くのウィンドウを同時に表示する必要が生じる。これにより、選択されたイベントとウィンドウ間のリンク関係の対応を取ることが困難になる。更に多くのウィンドウが表示されることにより画面が煩雑になり、ウィンドウ同士が重なり合う問題が発生する。

VisLink [16] は、複数の 2 次元パネルを用いることで、3 次元空間中での CMV を実現している。VisLink では、異なる 2 次元パネル上の関連情報間に明示的なリンクを張ることで、要素間の関係が認識しにくくなる問題を解決した。しかしながら、VisLink の手法は時間変化の可視化は考慮していない。また、複数イベントに関する頻度・詳細情報を表示するような大量の 2 次元パネルを用いる状況における可読性の評価は行っていない。

6. む す び

本論文では、人々の対象に対する行動及び興味を係り受け解析を用いてブログから抽出し、その時系列変化を 3 次元可視化し、探索するためのシステムを提案した。イベントデータベースと 3 次元可視化システムにより、ユーザが興味をもったキーワードに関する月ごとのトレンド変化観測が可能になった。また、長期にわたるイベントの頻度変化を可視化することで、イベントが流行しているタイミングや流行の周期性を発見することが可能になった。更に、それぞれのイベントに対して対話的に対象を掘り下げて可視化することで、「誰が、どこで、何を」などの詳細を探索可能にした。また、複数の事象に対する人々の行動・興味の共通点や違いをインタラクティブに比較することが可能になった。

本論文で提案したシステムは、様々なデータに対する時系列トレンド探索の為のアプリケーション構築に適応可能である。既に、動詞の代わりに形容詞を用いたイベントデータベースを構築することにより、ブログ空間における意見や感想の時系列変化を探索するアプリケーションも構築済みである。また、Twitter のような他の情報リソースに対する比較インタフェースも構築済みである。

現状の提案システムでは、探索に適切なキーワード

を選択するのがユーザにとって困難な場合がある．今後の課題としては，時期ごとの流行語リストから探索を開始するような仕組みや，より大量のキーワード（例えば，全議員名）を入力として与え全体像を俯瞰してからより詳細な探索を開始する仕組みが必要であると考えている．また，システムが自動的に特徴的な変化をしている領域を見つけそこへと誘導する仕組みも求められる．

本研究では個別のイベントを分析対象とした可視化システムを提案したが，図 3 の例のような（複文に含まれる）イベント間の依存関係（インフルエンザで死亡する → 不安が広がる）も分析対象として有益と考えられる．今後の課題として，動詞から動詞への係り受け DB を別途用意することで提案システムをイベント間関係の分析に応用することを検討したい．

謝辞 本研究は，文部科学省による「Web 社会分析基盤ソフトウェアの研究開発委託事業：多メディア Web 解析基盤の構築及び社会分析ソフトウェアの開発」からの支援を受けて行われた．よってここに謝意を表します．本論文の執筆にあたり，二人の査読者から貴重な助言・コメントを頂きました．ここに記して感謝の意を表します．

文 献

- [1] 喜連川優，豊田正史，田村孝之，鍛冶伸裕，“ボーンデジタル時代におけるウェブアーカイブとその活用基盤としての Socio-Sense,” 情報の科学と技術, vol.58, no.8, pp.372–375, 2008.
- [2] kizasi.jp. <http://kizasi.jp/>
- [3] BlogScope. <http://www.blogscope.net/>
- [4] BlogPulse. <http://www.blogpulse.com/>
- [5] 奥村 学，南野朋之，藤本稔明，鈴木泰裕，“blog ページの自動収集と監視に基づくテキストマイニング,” 第 6 回人工知能学会セマンティック Web とオントロジー研究会, 2004.
- [6] 颯々野学，“日本語係り受け解析の線形時間アルゴリズム,” 自然言語処理, vol.14, no.1, pp.3–18, 2007.
- [7] N. Yoshinaga and M. Kitsuregawa, “Polynomial to Linear: efficient Classification with Conjunctive Features,” Proc. 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009), pp.1542–1551, 2009.
- [8] 黒橋禎夫，長尾 眞，“京都大学テキストコーパス・プロジェクト,” 言語処理学会第 3 回年次大会, pp.115–118, 1997. <http://nlp.ist.i.kyoto-u.ac.jp/index.php?京都大学テキストコーパス>
- [9] M. Iwatate, M. Asahara, and Y. Matsumoto, “Japanese Dependency Parsing Using a Tournament Model,” Proc. 22nd International Conference on Computational Linguistics (COLING 2008), pp.361–368, 2008.
- [10] 山本悠二，増山 繁，“日本語係り受け解析における誤りの類型化と文構造の曖昧性について,” 言語処理学会第 15 回年次大会, pp.789–792, 2009.
- [11] G. Jacobson, “Space-efficient static trees and graphs,” Proc. 39th Annual Symposium on Foundations of Computer Science (FOCS 1989), pp.549–554, 1989.
- [12] M. Itoh, M. Toyoda, and M. Kitsuregawa, “An interactive visualization framework for time-series of web graphs in a 3D environment,” Proc. 14th International Conference on Information Visualisation (IV 2010), pp.54–60, 2010.
- [13] M. Itoh, “3D techniques for visualizing users’ activities on microblogs,” Proc. IET International Conference on Frontier Computing 2010, pp.384–389, 2010.
- [14] Y. Okada and Y. Tanaka, “IntelligentBox: A constructive visual software development system for interactive 3D graphic applications,” Proc. Computer Animation ’95 (CA ’95), pp.114–125, 1995.
- [15] Ed H. Chi, J. Pitkow, J. Mackinlay, P. Pirolli, R. Gossweiler, and Stuart K. Card, “Visualizing the evolution of web ecologies,” Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI ’98), pp.400–407, 1998.
- [16] C. Collins and S. Carpendale, “VisLink: Revealing relationships amongst visualizations,” IEEE Trans. Vis. Comput. Graphics, vol.13, no.6, pp.1192–1199, 2007.
- [17] S. Diehl and C. Görg, “Graphs, they are changing,” Proc. 10th International Symposium on Graph Drawing (GD 2002), pp.23–30, 2002.
- [18] T.M.J. Fruchterman and E.M. Reingold, “Graph drawing by force-directed placement,” Software — Practice & Experience, vol.21, no.11, pp.1129–1164, 1991.
- [19] 池田大介，南野朋之，奥村 学，“blog の著者の性別推定,” 言語処理学会第 12 回年次大会, 2006.
- [20] A. Kehoe and M. Gee, “Weaving web data into a diachronic corpus patchwork,” Language and Computers, vol.69, no.1, pp.255–279, 2009.
- [21] P. Resnik and A. Elkins, “The Linguist’s search engine: An overview,” Proc. 43rd Annual Meeting of the Association for Computational Linguistics, Interactive Poster and Demonstration sessions (ACLdemo ’05), pp.33–36, 2005.
- [22] M. Atterer and H. Schütze, “An inverted index for storing and retrieving grammatical dependencies,” Proc. 6th International Conference on Language Resources and Evaluation (LREC 2008), pp.62–65, 2008.
- [23] K. Inui, S. Abe, K. Hara, H. Morita, C. Sao, M. Eguchi, A. Sumida, K. Murakami, and S. Matsuyoshi, “Experience mining: Building a large-scale database of personal experiences and opinions from web doc-

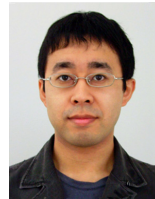
- uments,” Proc. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2008), pp.314–321, 2008.
- [24] M. Toyoda and M. Kitsuregawa, “A system for visualizing and analyzing the evolution of the web with a time series of graphs,” Proc. HYPERTEXT 2005, pp.151–160, 2005.
- [25] N. Nakazono, K. Misue, and J. Tanaka, “NeL2: Network drawing tool for handling layered structured network diagram,” Proc. 2006 Asia-Pacific Symposium on Information Visualisation (APVis 2006), pp.109–115, 2006.
- [26] G.G. Robertson, S.K. Card, and J.D. Mackinlay, “Information visualization using 3D interactive animation,” Commun. ACM, vol.36, pp.57–71, 1993.
- [27] G.G. Robertson, “Use of Animation in Information Visualization,” <http://research.microsoft.com/enus/events/fs2010/agenda.aspx>
- [28] Ed H. Chi and Stuart K. Card, “Sensemaking of evolving web sites using visualization spreadsheets,” Proc. 1999 IEEE Symposium on Information Visualization (InfoVis '99), pp.18–25, 1999.
- [29] U. Brandes and S.R. Corman, “Visual unrolling of network evolution and the analysis of dynamic discourse,” Information Visualization, vol.2, no.1, pp.40–50, 2003.
- [30] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, “ThemeRiver: Visualizing thematic changes in large document collections,” IEEE Trans. Vis. Comput. Graphics, vol.8, no.1, pp.9–20, 2002.
- [31] F. Wei, S. Liu, Y. Song, S. Pan, M.X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang, “TIARA: A visual exploratory text analytic system,” Proc. 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010), pp.153–162, 2010.
- [32] H. Notsu, Y. Okada, M. Akaishi, and K. Nijima, “Time-tunnel: Visual analysis tool for time-series numerical data and its extension toward parallel coordinates,” Proc. 2005 International Conference on Computer Graphics, Imaging and Visualization (CGIV 2005), pp.167–172, 2005.
- [33] 赤石美奈, 岡田義広, 中谷広正, 伊東幸宏, 田村貞雄, “史料の管理・検索・可視化機能を持つ歴史学研究支援統合環境の構築,” 情報学論, vol.40, no.3, pp.831–839, 1999.
- [34] J.C. Roberts, “State of the Art: Coordinated & multiple views in exploratory visualization,” Proc. Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007), pp.61–71, 2007.

(平成 23 年 11 月 14 日受付, 24 年 3 月 9 日再受付)



伊藤 正彦

2007 北大大学院情報科学研究科博士後期課程了。博士 (情報科学)。同大学大学院博士研究員を経て、現在、東大・生産技術研究所助教。情報可視化、ユーザインタフェースの研究に従事。日本データベース学会、情報処理学会各会員。



吉永 直樹

2005 東大大学院情報理工学系研究科博士課程了。2002 より 2008 まで日本学術振興会特別研究員 (DC1, PD)。2008 東大・生産技術研究所特任助教。博士 (情報理工学)。計算言語学・機械学習の研究に従事。



豊田 正史

1994 東工大・理・情報科学卒。1996 同大学院情報理工学研究科修士課程了。1999 同大学院情報理工学研究科博士後期課程了。博士 (理学)。同年、科学技術振興事業団計算科学技術研究員。2001 東大・生産技術研究所学術研究支援員、同大学同研究所産学官連携研究員、同大学生産技術研究所特任助教授、助教授を経て現在、同大学生産技術研究所准教授。ウェブマイニング、ユーザインタフェース、ビジュアルプログラミングに興味をもつ。ACM, IEEE CS, 情報処理学会、日本ソフトウェア科学会各会員。



喜連川 優 (正員:フェロー)

1978 東大・工・電子卒。1983 同大学院工学系研究科情報工学専攻博士課程了。工博。同年同大学生産技術研究所講師。現在、同教授。2003 同所戦略情報融合国際研究センター長。データベース工学、並列処理、Web マイニングに関する研究に従事。2009 ACM SIGMOD Edgar F. Codd Innovations Award 受賞。現在、本会副会長、日本データベース学会理事、情報処理学会フェロー、SNIA-Japan 顧問、本会データ工学研究専門委員会委員長 (1997~1998)、ACM SIGMOD Japan Chapter Chair (1999~2002) 歴任。VLDB Trustee (1997~2002)、IEEE ICDE, PAKDD, WAIM 等ステアリング委員、IEEE ICDE Program Co-chair (1999)、General Co-chair (2005)。