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8回)
SIG-AM-08-01

化学反応経路ネットワーク探索のための ビジュアルインタラクティビティのデザイン

The Visual Interactivity Design of the Chemical Reaction Map in RMapViewer

中小路 久美代 [1,2]　小田 朋宏 [2]　佐藤 寛子 [3]

Kumiyo Nakakoji[1,2], Tomohiro Oda[2], and Hiroko Satoh[3]

[1] 京都大学
[1]Kyoto University
[2] 株式会社 SRA
[2] Software Research Associates Inc.
[3] 国立情報学研究所
[3] National Institute for Informatics

Abstract: RMapViewer (Reaction Map Viewer) is an interactive visual environment for exploring a large network of chemical reaction paths, called the reaction map. The chemical reaction map consists of a set of atomic isomorphic 3-D structural configuration states that are theoretically derived by using the quantum mechanics. Each node is associated with its theoretically derived potential energy, and is connected with one or more other nodes via a transition link. The difference between the energy of two linked nodes represents necessary energy to transform the structural state on the one end to the adjacent state on the other end. The reaction map has been designed to support a user (i.e., a chemist) in exploring the map in a variety of ways, such as to find out possible molecule structures within a certain number of transitional steps from a focused node, or compare the energy values of possible transition paths between given two nodes. This paper describes RMapViewer by illustrating how a user interacts with the system through the different types of visualizations of the reaction map, sorting schemes, and transitional animations along with the model of a user's cognitive process in exploring the reaction map.

1. はじめに

RMapViewer（Reaction Map Viewer）[2]は、化学反応経路ネットワークを探索するためのインタラクティブな可視化環境である。化学反応経路のネットワークは、低分子を構成する原子の理論上あり得る幾何学的な構造をノードとし、それぞれのノード間での遷移可能性をリンクとする巨大ネットワークである。本プロジェクトでは、このネットワークを化学反応経路マップ（Reaction Map）と呼ぶ．ひとつの Reaction Map は、同じ原子の種類と個数で構成される，多数の異なる分子構造（異性体と呼ばれる）のノードから構成される。

RMapViewerは､分子デザインに携わる化学者が、着目するノードから指定するステップ数内で遷移可能性の吟味、二つのノード間の遷移経路の同定と比較といった探索タスクを行うことを目的として、可視化状態の遷移アニメーションと多視点並べ替えを擁している。本論では､RMapViewer の､ReactionMap 表示部分に焦点をあて、ReactinoMap 表示部分が提供するインタラクティティを列挙し、化学反応経路のデザインに携わる化学者の思考過程とビジュアルインタラクティビティの関係を考察する。

2. 化学反応経路ネットワーク探索タスク

分子は、同じ組成式でも構造が異なると異なる性質を有することが知られている。既知の化合物種の数は、年間数十万から百万種のオーダーで増加しており、これより遥かに多くの化合物種が存在可能と考えられている。組成を同じくする分子にあり得る幾何学的な構造についての研究は、化学情報学の分野において行われている[1]。分子構造をグラフとして数学的なグラフ処理によりトポロジカルな組み合

わせにより分子構造を発生させる方法などが一般的に知られている。本論で取り上げる RMapViewer は、分子のポテンシャルエネルギー局面を探索する GRRM という手法に基づくものである[3,4]。量子化学に基づいて理論的に存在可能な分子を創出し、電子状態や反応機構に関する情報を得ることが出来る。炭素原子 6 個と水素原子 6 個（C6H6）から成る分子の場合、GRRM（Global Reaction Route Maps）を用いると 4000 種以上の構造を生成できると考えられている。

ReactionMap の各ノードは、GRRM により求められた、複数の分子構造（異性体と呼ばれる）の幾何学情報（原子の種類と三次元座標値）と、そのポテンシャルエネルギー等の物理化学的パラメータから成る。ReactionMap を構成するノードには，EQ、TS、DC の 3 種類がある。EQ（Equibrium Structure：平衡構造）は、そのポテンシャルエネルギーが極小値にあり、安定的な分子構造を持つと考えられるものである。TS（Transitino State：遷移状態）は、二つの EQ を結ぶ化学反応経路上の最もポテンシャルエネルギーの高い分子構造である。DC（Dissociation Channel：乖離チャンネル）は、一つの分子構造が分解し、2 個以上の分子構造へと乖離した状態である。

ある EQ から、それにつながる TS を経て別の EQ につながる経路が、化学反応の１ステップである。ある EQ から別の EQ まで、リンクを辿ったパスが、化学反応経路となる。途中に、他の EQ が 1 個以上含まれる場合もあり、その個数に応じて、化学反応ステップが増えることになる。リンクは無向であり、ES1 から ES2 への変化（反応）には TS を超えるためのエネルギーが必要となる。すなわち、ES1、ES2、TS のポテンシャルエネルギーの値をそれぞれ e(ES1), e(ES2), e(TS) とすると、e(ES1)<e(TS), e(ES2)<e(TS)であり、ES1 から ES2 への変化には、e(TS)-e(ES1)の遷移エネルギーが、ES2 から ES1 への変化には e(TS0-e(ES2)の遷移エネルギーが必要と考えられる。

化学反応経路を解析する際の基本要件は、佐藤ほか[2]に詳しいが、ここではその概要を述べる。

注目する属性としては、分子の構造と、ポテンシャルエネルギー、二つの EQ 間をつなぐパスの有無、パスの長さ、およびパスにおける遷移状態エネルギーである。

分子の構造には、平面構造（２次元構造）表現と立体構造（３次元構造）表現とがある。

ポテンシャルエネルギーは、分子の内部エネルギーを示す物理化学的数値であり、値が低ければ安定、高ければ反応性の高さを示す尺度となる。二つの EQ を結ぶ TS のポテンシャルエネルギーは、その二つの分子構造間で変化させるために必要なエネルギー値を示す尺度となる。

二つの EQ 間をつなぐパスの有無は、一方を反応物、他方を生成物とする場合の、反応可能性を示す。パスが存在する場合、パスの長さは、反応ステップ数を示す。遷移エネルギーのより低いパスを経るものが、化学反応を起こし易いと考えられる。

3. RMapViewer のデザイン

本章では、RMapViewer における、ReactionMap（反応経路）表示のためのビジュアルインタラクションのデザインについて述べる。

RMapViewer のビジュアルインタラクティビティは、RMapViewer のユーザとなる化学者へのインタビューに基づいてデザインした。ReactionMap のビジュアル表現にあたっては、下記の項目が必要なインタラクション要件として同定された。

(1) 反応系から生成系へのパスとして、場合によっては何十個、何百個というパスが、化学反応経路ライブラリから検索結果として得られることも考えられる。これらの検索結果をビジュアルに表示し、化学者がインタラクティブに選別していくためのユーザインタフェースが必要である。

(2) 反応系から生成系へのパスは、途中にある反応中間体（Intermediate State）と、それぞれの間に存在する遷移状態（Transition State）との列から構成される。安定状態の数としては、0 個から数十個までが想定される。

(3) 本システムを利用するユーザは、一日に十回から数十回という回数で、反応経路の探索を行うことになると考えられる。入力となる反応系と生成系のうち、どちらか片方がより重要といったことは必ずしも一意に定まらない。本システムを利用するユーザの目的によって異なる。合成を目的とするユーザにとっては生成系が重要であり、反応解析を目的とするユーザにとっては反応系が重要となる。

(4) 探索されてきた各パス毎に、以下の情報が提示される必要がある：

- 反応系（R: Reactant）のエネルギー値
- 生成系（P: Product）のエネルギー値
- ステップ数（パスに含まれる中間体（M: Intermediate State）の個数）
- 遷移状態（TS: Transition State）のエネルギー値が最大となる TS とそのエネルギー値（maximum

energy)
- 遷移状態（TS: Transition State）のエネルギー値が最小となる TS とそのエネルギー値（minimum energy)
- パス全体の最小エネルギーと最大エネルギーの差の値
- 隣り合う M と TS のエネルギー差が最大となる TS およびそのエネルギー値（最大活性化エネルギー: maximum activation energy）
- R, P, M のエネルギー値の中で最も低いエネルギー値（中間体の最小エネルギー値）

(5) R, P, M および TS のエネルギー値は、原則として kJ/mol 単位で表すものとする。エネルギー値として取り得る値は、0 から 3000kJ/mol 程度である。必要に応じて hartree 単位（Eh）も併記する。

(6) 検索結果として得られたパスは、ステップ数が n の場合、[R, TS(1), M(1), TS(2), M(2), ... TS(n-1), M(n-1), TS(n), P]で表される。M(x-1)から M(x)への間の TS(x)は、原則として最もエネルギー値の低いもののみの提示でよい。但し、ユーザの要求に応じて、より高いエネルギー値の TS についても調べられるようになっていることが望ましい。

(7) 検索されたパスは、[R, TS(1), M(1), TS(2), M(2), ... TS(n-1), M(n-1), TS(n), P]の各状態を、横軸は等間隔に、縦軸はそのエネルギー値でプロットしてビジュアルに表現したい。

(8) 複数のパスは重畳表現して比較できるものとしたい。各パスは、ビジュアルな表現のグラフ上で選択できるようにしたい。

これらの要件を基にして、RMapViewer の ReactionMap 表示部分のデザインは、一般的に知られる乗り換え案内ウェブサービスの経路探索のインタラクティビティをベースとすることとした。乗り換え案内では、出発駅と到着駅を指定し、その経路を探索する。表示した経路の候補は、それぞれの経路の所用時間や運賃といったプロパティで並べ替えすることが出来る。ReactionMap でも同様に、出発地点となる反応物と到着地点となる生成物とを指定し、そのパスが複数表示されることとした。パスを選択するとそのプロパティ値が表示され、その値を用いて並び替えを出来るようにした。図 1 に、スケッチとして作成した ReactionMap の表示部分を示す。RMapViewer の ReactionMap 表示部分のインタラクティビティは、このスケッチに基づいて実装されている。

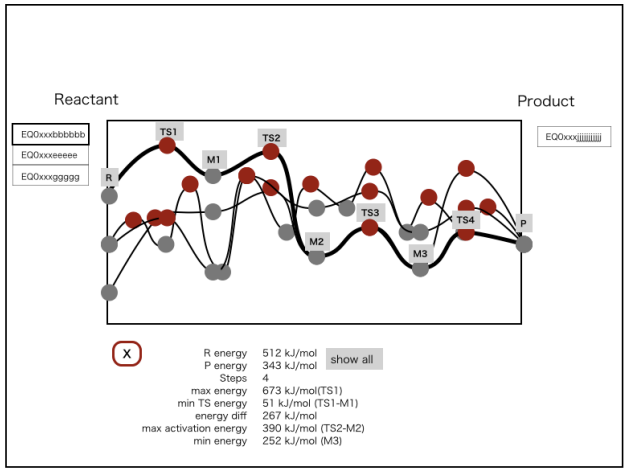


図 1: ReactionMap 部分のビジュアルインタラクティビティのデザインにおいて作成されたスケッチ

4. RMapViewer におけるビジュアルインタラクティビティ

本章では、RMapViewer の現状の実装における ReactionMap 表示部分（図 2）のビジュアルインタラクティビティについて説明する。

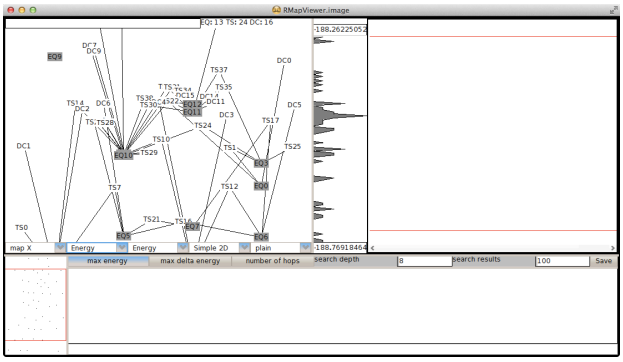


図 2: RMapViewer における ReactionMap 表示部分（CH2O2）

(1) ReactionMap 表示部分において、ノードは下記の種類で表現される：
- 分子 ID でのテキスト
- 分子の 3 次元構造を gif アニメーション
- 分子に関する属性情報の詳細（分子 ID、ポテンシャルエネルギー値、smiles 記法、inchi 記法、canost 基本）

(2) ReactionMap 表示の右側のウィンドウは、表示されている全分子のポテンシャルエネルギーの分布ヒ

ストグラムを表示する。表示されている全分子のポテンシャルエネルギー値の最小値（下）から最大値（上）までを縦軸とし、対応するポテンシャルエネルギー分子の数が横軸にプロットされている。表示部分の上辺を下にドラッグし、下辺を上にドラッグすることで、その領域内の値のポテンシャルエネルギーを有する分子のみを左の ReactionMap ウィンドウに表示する。

(3) ReactionMap のズームイン／アウトは、左下のフォーカスウィンドウ表示部分で行う。全体に対して表示されている部分が赤枠で囲われている。

(4) マップはウィンドウ左下を基点とする、x 方向（横軸）、y 軸（縦軸）、z 軸（奥行き方向）から成る 3 次元で構成される。マップウィンドウ下部のプルダウンメニューによって、それぞれの方向で、並べ方の軸を下記から選択することができる。

- Energy：その分子のポテンシャルエネルギー値の昇順で上から下に並べる
- Kind：EQ、TS、DC の別に並べる
- Hops：Reactant（反応物）指定時に、それからの到達ステップ数で並べ替える

なお、並べ替え時のビューの切り替えの際にはアニメーション表示を行い、表示コンテキストの連続性を保つ。図 2 は、y 軸を Energy 状態でソートした状態である。

(5) EQ を一つ選び（図 3 で EQ2）、メニューを介して Reactant（反応物）に設定することが出来る。ノードは一重線の青枠で囲まれる。それによって、右側のパス表示ウィンドウに、パスとそのポテンシャルエネルギー値を表示する。パス表示ウィンドウでは、横軸を反応ステップ、縦軸をポンテンシャルエネルギー値としてノードを表示する。ステップによって、同じ分子 ID が複数回表示されることもある。

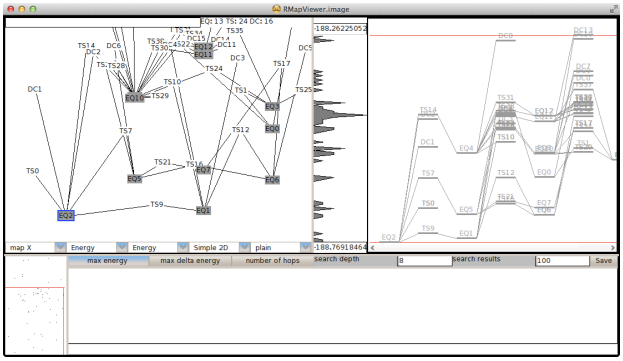


図 3: ReactionMap 表示部分における反応物（Reactant）の設定

(6) 別の EQ を選んで（図 4 では EQ1）、メニューを介して Product（生成物）に設定することが出来る。ノードは二重線の青枠で囲まれる。それによって、下記のパス詳細表示ウィンドウに、EQ2 から EQ1 へと至るすべてのパスと、それぞれのポテンシャルエネルギーに関する情報がリストとして表示される。パス検索時の深さを指定することが出来る。デフォルトの探索深さは 8 である。

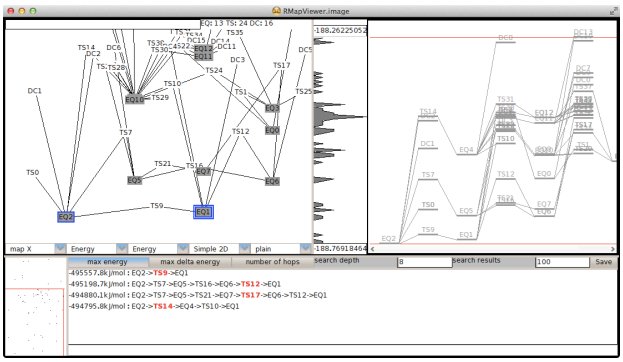


図 4: ReactionMap 表示部分における生成物（Product）の設定

(7) パス表示ウィンドウ内のリストの表示は、上のタブで以下に示す順に並べ替えることができる（図 5）：

max energy：パス内に含まれる分子がもつ最大エネルギーの値で昇順にリストをソートする。そのエネルギー値を有する分子は赤字で表示される。

mas delta energy：パス内に含まれる分子の、最大変化エネルギーの値（直前の分子のポテンシャルエネルギー値との差）でリストを昇順にソートする。そのエネルギー値を有する分子は赤字で表示される。

number of hops：パスの反応ステップ数（i.e., 含まれる分子数）でソートする。

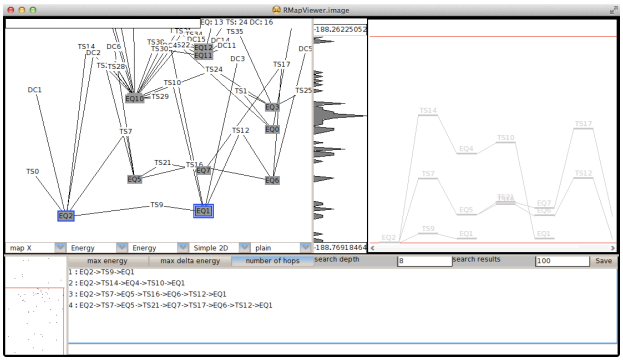


図 5：パス詳細ウィンドウのパスリストの並べ替え

(8) パス詳細表示ウィンドウに表示されているリストから、パスをひとつ選択すると、マップウィンド

ウおよびパス表示ウィンドウの中で、他のパスがグレイアウトされ、指定した EQ2 から EQ1 へのパスのみが協調表示された状態になる（図 6）。

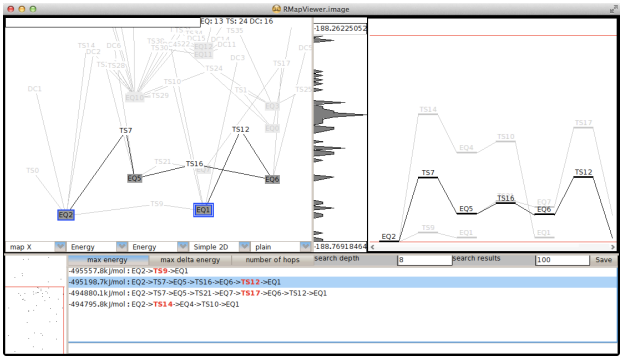


図 6：パス詳細ウィンドウにリストされたパスの一つを表示する

(9) パスを選択した状態で、右クリックにより Open in jmol コマンドを選択することで、反応物から生成物に至るまでのパス内の、各ステップ毎の分子の３次元構造の変化のアニメーションムービーを別ウィンドウで生成する（図 7）。

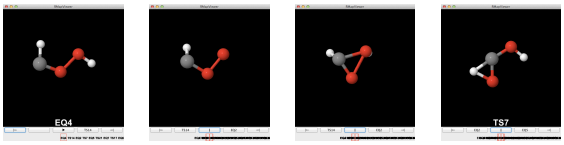


図 7：指定した反応物から生成物へのパスを辿る分子の３次元構造表現のアニメーションの生成

5. むすび

本論では、RMapViewer における ReactionMap のビジュアルインタラクティビティと、それがベースとしたデザインの考え方について解説を行った。ReactionMap におけるインタラクティビティの快適さは、表示および反応速度に大きく依存する。RMapViewer の開発にあたっては、プロトタイプを繰り返しながら、経路探索等の計算速度を踏まえつつ、必要なインタラクティビティの同定を行っている。

RMapViewer は現在も開発が進行中であり、着目すべきノード自体を検索する機能等を今後追加していく予定である。RMapViewer は、広く化学者に公開することを想定しており、多数の利用ケースを踏まえながら、ビジュアルインタラクティビティについても展開することを目指す。計算処理速度の高速化と頑健化というバックエンドの実装方式と、化学反応経路の設計に携わる化学者（ユーザ）の知識創造活動のためのフロントエンドとなるビジュアルインタラクティビティ[5]の実装方式との関係性について、今後研究を進めていきたいと考えている。

謝辞

本研究の一部は、JSPS 科研費挑戦的萌芽研究 25540017、国立情報学研究所共同研究費、情報・システム研究機構データ中心科学リサーチコモンズ基盤整備事業、および科学技術振興機構 CREST の助成による。

参考文献

[1] 佐藤寛子, 化学情報学：化学反応の系図と反応予測, 丸善, 2003.

[2] 佐藤寛子, 小田朋宏, 中小路久美代, 宇野毅明, 田中宏明, 岩田 覚, 大野公一, 「埋蔵分子」発掘プロジェクト：化学反応経路マップのインタラクティブ可視化に向けて, 情報処理学会, インタラクション 2014, インタラクティブ発表, March, 2014.

[3] 中小路久美代, 山本恭裕, 創造的情報創出のためのナレッジインタラクションデザイン, 人工知能学会論文誌, Vol.19, No.2, pp.154-165, March, 2004.

[4] Ohno, K.; Maeda, S. A Scaled Hypersphere Search Method for the Topography of Reaction Pathways on the Potential Energy Surface, Chemical Physics Letters, Vol. 384, pp277-282, 2004.

[5] Ohno, K.; Maeda, S. J. Phys. Chem. A 2006, 110, pp.8933-8941, 2006.

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8回)
SIG-AM-08-02

ユーザ間の関係可視化による コミュニケーション支援システムの提案

Proposal of Communication Support System by Visualizing Inter-user Relationship

鈴木友也[1*] 上村春貴[2] 高間康史[1]
Yuya Suzuki[1] Haruki Kamimura[2] Yasufumi Takama[1]

[1] 首都大学東京大学院システムデザイン研究科
[1] Graduate School of System Design, Tokyo Metropolitan University
[2] 首都大学東京システムデザイン学部
[2] Faculty of System Design, Tokyo Metropolitan University

Abstract: 本稿では，複数ユーザ間の関係を可視化するコミュニケーション支援システムを提案する．文書の共同執筆や企画立案などの共同作業では，他者との考えの共通・相違を認識し，全体の方針を決定するプロセスが必要となる．提案システムではこの作業を共同作業のためのコミュニケーションとして捉え，キーワードベースでユーザ間の関係を可視化することで支援する．配置の主体となるユーザをインタラクティブに変更可能とすることで，多様な視点からの検討を可能とする．

1 はじめに

本稿では，複数ユーザ間の関係を可視化するコミュニケーション支援システムを提案し，データ可視化のための JavaScript ライブラリである D3.js [1]を用いて実装したプロトタイプシステムを用いて行った予備実験結果を示す．

研究やビジネスにおいて，文書の共同執筆や企画立案などの共同作業を行うことが多い．また，共同作業を行う際，会議やディスカッションによるグループでのコミュニケーションにて，全体の方針を決定することも一般的である．角ら [1] は，コミュニケーションをとる目的として，以下の二点を挙げている．

1. 伝え手の心の中にある知覚・感情・思考を，過不足なく正確に受け手に伝達する．
2. 他人とのコミュニケーションをとることにより，それまではもやもやとしていたアイデアが明確に認識されたり，心の中に新たな発想が生じるといった効果を期待する．

また，ディスカッションによるコミュニケーションの際に直面する問題として，ディスカッションへの参加者の背景知識が似通っていたり，ディスカッションの文脈に捕らわれ過ぎることが原因で，参加者の視野が狭まり，アイデアが枯渇すること，ディスカッションの全体的な構造を把握していないがために，過去の議論を無視した発言がなされたり，同じ話題が繰り返されることがあることが挙げられている [2]．そこで，ディスカッションなどのコミュニケーションの場面において，他者との考えの共通・相違を認識し，全体の方針を決定するプロセスが必要と考える．

そこで本稿では，キーワードベースでユーザ間の関係を可視化し，多様な視点からの検討を行うことができる，コミュニケーション支援システムを提案する．提案するコミュニケーション支援システムは，各ユーザに関するメインキーワードと，それに付随する関連語をノードとして，各ノードをリンクするネットワーク形式で表示する．各ユーザのキーワード空間は，表示・非表示を選択することが可能である．複数のユーザのキーワード空間を表示する場合，ノードは各ユーザ毎に色分けされ表示される．また，本システムではメインユーザを指定することで，そのユーザのキーワードを中心として関係性を可視化することが可能である．これにより，多様な視点からの検討が可能となる．さらに，ノードとエッジの追加・削除，キーワードの役割や表示位置をインタラクティブに変更することができる．単に眺めるだけでなく，これらの編集作業を通じたコミュニケーションの支援効果も期待できる．

本稿では，提案システムの可視化方法やインタラクションについて述べると共に，構築したプロトタイプ

*連絡先： 首都大学東京大学院 システムデザイン研究科
〒 191-0065 東京都日野市旭ヶ丘 6-6
E-mail: suzuki-yuya1@ed.tmu.ac.jp

[1] http://d3js.org

システムを用いて行った予備実験結果について示す.

2 関連研究

2.1 人物間の関係可視化

近年，人のつながりや社会性が情報技術分野でも注目されており，社会ネットワーク分析と情報技術をつなぐ様々な研究が行われている [3]．その中でも Web 上の情報を用いて人物間の関係可視化を提案している研究が多い.

松尾ら [4] は，Web 上の情報のみを用いて特定のコミュニティの人間関係を抽出する手法を提案している．提案手法では，人工知能学会のコミュニティを対象に，検索エンジンを利用して特定の 2 人の人物に関する文書を検索し，その内容やヒット件数から 2 人の関係の強さや関係の種類を判断する．2 人の関係の強さには Simpson 係数を用いている．関係の種類は,「共著関係」,「同研究室関係」,「同プロジェクト関係」,「同発表関係」の 4 種類としている．関係の種類を抽出するために 2 人の名前を AND 検索し，上位 5 ページを取得する．その 5 ページから，著者らが定めた 6 つの属性を抽出し，その属性の値から判別ルールによって関係を決定する．この方法により，適合率・再現率ともに高い値で著者関係を抽出可能であることを示している.

寺川ら [5] は，実社会のコミュニティと比較すると，ソーシャルメディアのコミュニティは，ユーザ間の相互関係などの潜在的な要素の把握が難しいと考え，ユーザの性格診断に基づき，ユーザ間の潜在的な関係を可視化する手法を提案している.

金ら [3] は，従来の人間関係ネットワークでは弱いとされていた人物間の関係であっても，相対的に強い関係にある人々を見つける方法を提案している．提案手法では，絶対的ルールと相対的ルールを組み合わせネットワークを抽出している．絶対的ルールでは，共起の値が一定の閾値以上になるノードをエッジで繋ぐ．相対的ルールでは，絶対的ルールで繋がらない様なネットワーク全体では共起が低い人物関係でも，各ノードから見て共起が高い場合は M 人までエッジを繋ぐ.

人物間の関係可視化では，Web 上の情報を用いて人物間の関係可視化するシステムの研究が多いが，本稿では，グループ活動における人物間の関係を可視化することを目的とする点で異なる.

2.2 共同作業におけるコミュニケーション支援

コンピュータによる共同作業支援は CSCW(Computer Supported Cooperative Work) と呼ばれ，数多くの研究がなされている.

由井薗ら [6] は，共同作業における発想支援を目的とした KJ 法支援システムである KUSANAGI を提案している．KJ 法は，1960 年代後半頃より日本国内で普及し始めた会議技法であり，意見出し，グループ編成，グループ関係の図解化，文章化の 4 段階の作業に分けられる．KUSANAGI は意見出しやグループ編成の作業などを，複数台の計算機とモニタを用いて行うことができる.

小宮ら [7] は，複数人がイメージを共有して創作を行う形態の共同作業を「共創型共同作業」と呼び，ユーザの持つ作品イメージを可視化することで，複数ユーザ間でのイメージ共有を支援するシステムである MochiFlash を提案している．MochiFlash では，ユーザが特定の作業に対して持っているイメージを言葉や画像などで表したものを「イメージマップ」と呼び，このイメージマップを 2 次元空間上に配置し，各イメージマップのリンク化・グループ化を行うことができる．このシステムを用いることにより，ユーザが持つイメージを外在化し，複数ユーザ間での差異を明確にすることや，イメージの変遷を明確に捉えることができるとしている.

渡辺ら [8] は，共同作業の場では共有したい情報としたくない情報があるとし，共同作業領域と個人作業領域を同時に確保したテーブル型のグループウェアシステムを提案している．提案しているグループウェアシステムは，個人作業用のディスプレイの上に，透過性のディスプレイを重ね，これを共同作業領域としている．渡辺らは，作業領域を 2 重化することによる利点として，下記の 3 点を挙げている.

1. 共同作業領域を，テーブル上面の最大領域まで広げられる
2. 個人作業領域にソフトウェアキーボードなどの個人ツールを表示することで，共同作業領域を遮ることなく作業できる
3. 個人作業領域に共同作業領域の操作面を用意することで，共同作業領域の任意の点を遠隔操作できる

白井ら [9] は，複数タブレットを用いた複数人でのデータ分析作業を支援するシステムを提案している．提案システムでは，タブレット上に 1 つのビューを表示し，他のタブレットと共通のビュー・異なるビューを見ることができる．また，各タブレットの画面を連動させ，ビューに操作を加えた場合，他のタブレットにも反映させることで，他者がどのような点に着目したのかを把握することを可能にしている.

3 提案するコミュニケーション支援システム

3.1 システム構成

推薦システムは，会議などのコミュニケーションの場面での利用を対象とし，Web アプリケーションとして構築する．

提案するコミュニケーション支援システムのシステム構成を図 1 に示す．提案システムは，各ユーザのキーワードなどの情報を格納する RDB および，RDB からキーワードなどの情報を取り出し可視化するアプリケーションサーバから構成される．提案システムは用途を限定せず様々に利用できることを想定して構築している．そのため，初期メインキーワード・関連語は，ユーザが自ら選んだワードを RDB に登録する方法と，キーワード・関連語抽出モジュールを用い，抽出結果を RDB に登録する方法の 2 種類を想定している．ユーザは Web ブラウザを通じて可視化されたキーワードの閲覧・操作を行う．また，ユーザが操作した結果・ログは RDB に登録される．

実装に関して，RDB には PostgreSQL [2]を，Web アプリケーションフレームワークには Ruby on Rails [3]を用いた．RDB は，KeywordTBL・UserTBL・Relation・Owner・Log の 5 つのテーブルからなる．KeywordTBL テーブルと UserTBL テーブルには，それぞれキーワードとユーザを登録する．Relation テーブルには，リンクする 2 つのキーワードとその関連度を保持する．また，キーワード間の関係はユーザ毎に与えられるため，対応ユーザの ID もこのテーブルに保持する．Owner テーブルには，キーワード毎に対応ユーザの ID および役割（メインキーワード/関連語）を登録する．また，Owner テーブルにはキーワードの操作結果も登録する．Log テーブルには，ログ解析のためのユーザ操作ログを登録する．

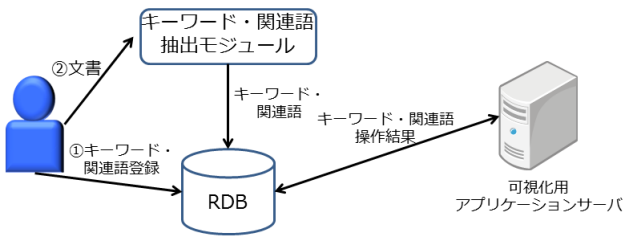


図 1: システム構成図

[2]http://www.postgresql.org
[3]http://rubyonrails.org

3.2 キーワード・関連語登録

3.1 節でも述べた通り，RDB へのキーワード登録には以下の 2 パターンを想定する．

1. ユーザが直接 RDB に登録
2. キーワード・関連語抽出モジュールを用い，抽出結果を RDB に登録

パターン 2 のキーワード・関連語抽出には，メインとなるキーワードとその関連語に分類してキーワード抽出を行う既存手法が利用可能である．KeyGraph [10]は，文書の主張内容を表すキーワードの抽出を行う手法であり，「土台の形成」・「屋根の形成」・「キーワードの抽出」の 3 フェーズからなる．土台の形成では，文書中の単語を出現回数でソート後，上位 M 語を抽出し，文書から生成した共起グラフを用いて共起度の高い語をリンクし土台とする．屋根の形成では，土台から文書の主張を表す語を屋根として取り出す．キーワード抽出では，屋根の語と土台の語のリンクの強さでソート後，上位 12 語をキーワードとして抽出している．

展望台システム [11] では，文書から重要文を抽出して要約を生成するために，文章からの特徴キーワード・周辺キーワードを抽出する手法を提案している．文章から形態素解析器 (Chasen [4]) によって名詞・動詞・形容詞をキーワード候補として抽出した後，文章中での出現頻度に基づき，周辺キーワードを最大 20 語取得する．周辺キーワードと同時に出現しやすい単語を中心キーワードとして最大 5 語抽出する．特徴キーワードは，中心キーワードと同時にしか出現しない単語としている．

岡田ら [12] は，ウェブ上のデータからキーワードを自動的に抽出する手法を提案している．提案手法では，ウェブ上のテキストデータをパラグラフ単位に分割し，各パラグラフ内で出現頻度の高い単語をキーワード候補とする．次に，話題の転換部を探して話題ごとのキーワード候補を得た後，文書を単位とした IDF(Inverse Document Frequency) と話題を単位とした IDF を用いて全体のキーワードと話題のキーワードを抽出する．

キーワード・関連語抽出モジュールにこれらのアルゴリズムなどを採用することで，ユーザが持つ文書からメインキーワード・関連語を抽出し，RDB へ登録する．

3.3 インタフェース・機能

図 2 に構築したコミュニケーション支援システムのインタフェースを示す．システムはキーワード表示部とキーワード操作部からなる．

[4]http://chasen.naist.jp

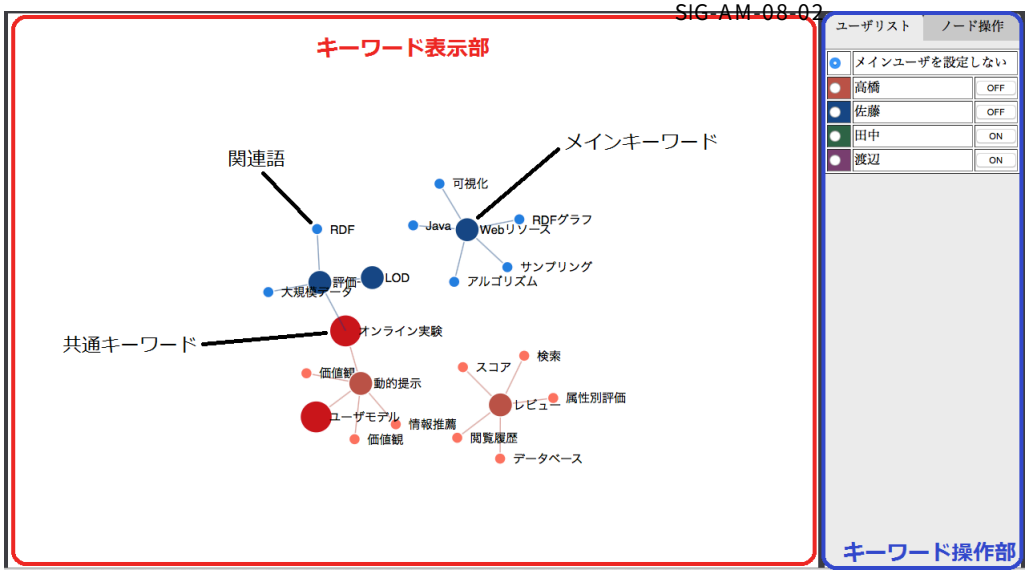


図 2: コミュニケーション支援システムのインタフェース

キーワード表示部では，選択した複数ユーザのキーワードをノードとしたネットワークを表示する．ノードは，各ユーザに割り当てられた色で表示される．各ユーザのメインキーワードとなるノードは大きく表示される．各ユーザについて，メインキーワードと関連語はリンクで接続されるためユーザ毎にいくつかのサブネットワークを構成するが，ユーザ間で共通するキーワードが存在する場合にはサブネットワークが接続される．他のユーザとの共通のキーワードは赤色の大きなノードとして表示される．キーワードの表示方法は D3.js の視覚化コンポーネントのひとつである Force-Directed Graph を用いる．

コミュニケーションには，各話題の中心となる人物がいると考える．そこで提案システムでは，話題の中心となる人物にあたるメインユーザを設定する機能を備える．メインユーザを指定することにより，コミュニケーションにおける話題の中心が把握しやすくなると考える．

キーワード操作部は，ユーザリストタブとノード操作タブからなる．ユーザリストタブは各ユーザのキーワードの表示・非表示の選択，メインユーザの設定を行う．ノード操作タブのインタフェースを図 3 に示す．ノード操作タブでは，ユーザリストタブで選択されたメインユーザに関する操作やノード・エッジに関する操作を行う．ノード操作タブで閲覧できる情報は以下の 2 つである．

1. 現在のメインユーザ
2. 選択したノードの情報

また，行える操作は以下の 4 つである．

1. メインユーザにノード追加
2. ノードの役割変更
3. ノード削除
4. メインユーザのノードと他ユーザのノード間のエッジ追加

キーワード表示部の任意のノードをダブルクリックすることで「選択したノードの情報」にノードの所有者・キーワード・役割が表示される．また,「このキーワードを追跡する」のチェックボックスを On にすると，選択中のメインユーザにとって重要な単語と判断される．メインユーザのサブネットワークに，他ユーザのノードの追加や，任意のノードの役割の変更・削除を行える．また，選択したメインユーザの任意のノードと他のノード間にリンクを追加することができる．このように，表示するユーザの選択や，メインユーザを変えノードやリンクを操作することによって，多様な視点から検討・分析が行える．

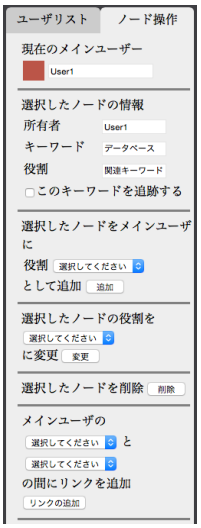


図 3: ノード操作タブのインタフェース

4 コミュニケーション支援に関する予備実験

4.1 実験概要

3.1 節でも述べた通り，提案システムは，会議などのコミュニケーションの場面での利用を想定しているため，複数人で実験を行った．

本実験では，同じ研究室に所属する 20 代の工学系大学院生 4 名に，研究室の研究内容を説明するポスターを協力して作成する前の打ち合わせに，提案システムを利用してもらった．4 名の実験協力者には，事前に自身の研究内容を表すメインキーワード 2 語と，各メインキーワードに関係する関連語を 5 語ずつ選んでもらい，直接 RDB に登録した．実験協力者 4 名が選んだメインキーワードと関連語を表 1 に示す．ユーザ間で共通するキーワードは，価値観とビッグデータの 2 語であった．実験前に提案システムの使用方法を説明し，実験時間を 1 時間程度と定め，システムを利用してもらった．実験後に，実験協力者にアンケートに回答してもらった．

表 1: 各実験協力者が選んだメインキーワード・関連語

実験協力者	メインキーワード	関連語
協力者 1	レビュー	閲覧履歴
		スコア
		データベース
		属性別評価
		動的提示
	情報推薦	**価値観**
		ユーザモデル
		オンライン実験
		ホテル
		アンケート
協力者 2	クラスタリング	分類
		ビッグデータ
		データマイニング
		階層
		集合
	ツリーマップ	表示
		木構造
		長方形
		色
		大きさ
協力者 3	LOD	Web of Data
		ビッグデータ
		RDF
		Web リソース
		探索的 LOD 分析
	可視化	可視化システム
		サンプリング
		RDF グラフ
		データ構造
		ネットワークグラフ
協力者 4	推薦システム	アイテムモデリング
		データ分析
		映画
		ユーザレビュー
		推薦理由
		ユーザレビュー
		属性
	価値観	極性一致
		評価一致
		相関ルール

4.2 実験結果

図 4 に実験終了時のネットワーク配置を，図 5 に完成したポスターを示す．2 つの図を比較すると，図 4 の赤枠・青枠内のネットワークと図 5 の赤枠・青枠の内容がそれぞれ同じ色の枠同士で対応していることがわかる．また，アンケートにて，共通した部分のキーワードが可視化されていることで，ポスターの概要を決めるのに参考になったとの回答を得た．これは，実験中に提案システムのキーワード表示部をポスターとして見立て，キーワードの繋がりを見ながら議論を行っていたためであると考える．

また，実験協力者によるアンケートでの評価では，4 名中 3 名が提案システムはポスター制作に役立ったと回答していた．役立った理由として，互いに関連しているキーワードを発見することができ，ポスターの全体像をまず決めてからから個別の話題に入ることができたとしている．また，役立たなかった理由として，他者の特徴的なキーワードと自分のキーワードを比較できる点は良いが，口頭やメモで十分補えるとしている．

提案システムの UI の問題点として，ノードやエッジの操作後，ブラウザを更新しなければいけない点，Force-Directed Graph を用いてたため，ノードの移動に慣性が残りすぎ，動作が遅く感じる点などが挙げられた．追加してほしい機能としては，文字やノードの大きさの変更，初期状態では存在しないキーワードの追加などが挙げられた．UI の問題点，追加してほしい機能は，今後提案システムの改良の際に考慮する予定である．

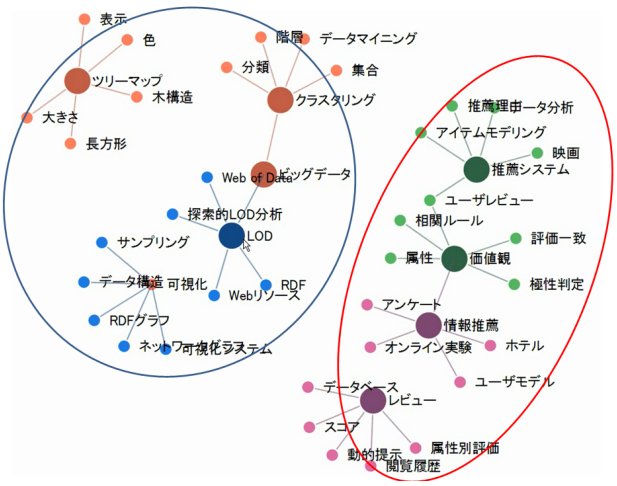


図 4: 実験終了時のネットワーク配置

図 5: 完成したポスター

5 まとめ

本稿では，複数ユーザ間の関係を可視化するコミュニケーション支援システムを提案し，提案システムの可視化方法やインタラクションについて述べると共に，構築したプロトタイプシステムを用いて行った予備実験結果について示した．

提案システムは，各ユーザに関するメインキーワードと，それに付随する関連語をノードとして，各ノードをリンクするネットワーク形式で表示する．また，ノードとエッジの追加・削除，キーワードの役割や表示位置をインタラクティブに変更することができる．提案システムを用いて，複数人によるポスターの協調作成に関する予備実験を行った結果，提案システムに表示されるネットワークをポスターのレイアウトと見立てることで，ポスター作成に有効なコミュニケーションが行われたことを示した．

今後は，UI の改善，文字やノードの大きさの変更・初期状態では存在しないキーワードの追加などの機能追加，追跡ワード・削除ワードの活用をすることで，更にコミュニケーションを支援できるシステムの構築を目指す．

参考文献

[1] 角 康之, 小川 竜太, 堀 浩一, 大須賀 節雄, 間瀬, 健二: 思考空間の可視化によるコミュニケーション支援手法, 電子情報通信学会論文誌 *A*, Vol.79, No.2, pp.251-260 (1996)

[2] 角 康之, 西本 一志, 間瀬 健二: グループディスカッションにおける話題空間の可視化と発言エージェント, 情報処理学会研究報告. 情報学基礎研究会報告, Vol.96, No.88, pp.103-108 (1996)

[3] 金 英子, 松尾 豊, 石塚 満: Web 上の情報を用いた弱い社会的関係のネットワーク抽出手法 (データマイニング), 電子情報通信学会論文誌 *D*, Vol.91, No.3, pp.709-722 (2008)

[4] 松尾 豊, 友部 博教, 橋田 浩一, 中島 秀之, 石塚 満: Web 上の情報からの人間関係ネットワークの抽出, 人工知能学会論文誌, Vol.20, No.1, pp.46-56 (2005)

[5] 寺川 晃司, 浦 正広, 中 貴俊, 山田 雅之, 遠藤 守, 宮崎 慎也: ソーシャルメディアにおけるユーザ間の潜在的関係の可視化手法の提案, 情報処理学会研究報告. *EC*, Vol.2011, No.14, pp.1-2 (2011)

[6] 由井薗 隆也, 宗森 純, 重信 智宏: 大画面共同作業インタフェースを持つ発想支援グループウェア KUSANAGI が数百データのグループ化作業に及ぼす効果, 情報処理学会論文誌, Vol.49, No.7, pp.2574-2588 (2008)

[7] 小宮 香織, 関口 佳恵, 庄司 裕子, 加藤, 俊一: 共創型共同作業のための合意形成支援システム:MochiFlash, 感性工学研究論文集, Vol.7, No.4, pp.675-684 (2008)

[8] 渡辺 亮太, 松浦 吉祐, 郷 健太郎: 共同作業領域と個人作業領域を同時に確保したテーブル型グループウェア, 情報処理学会全国大会講演論文集, Vol.7, No.4, pp“4-415”-“4-416” (2008)

[9] 白井 智子, 萬成 亮太, 三末 和男, 田中 二郎: 複数タブレットを用いた共同分析作業のための視覚的表現および操作の検討, 情報処理学会研究報告. *HCI*, Vol.2013, No.15, pp.1-8 (2013)

[10] 大澤 幸生, ベンソン ネルス E, 谷内田, 正彦: KeyGraph : 語の共起グラフの分割・統合によるキーワード抽出, 電子情報通信学会論文誌. *D-I*, Vol.82, No.2, pp.391-400 (1999)

[11] 砂山 渡, 谷内田 正彦: 文章要約のための特徴キーワードの発見による重要文抽出法 : 展望台システム, 情報処理学会研究報告. 自然言語処理研究会報告, Vol.2000, No.11, pp.103-110 (2000)

[12] 岡田 真, 浜田 浩史, 宝珍 輝尚: マルチメディアデータの効率的検索のためのキーワード自動抽出手法 (検索とキーワード・概念抽出), 情報処理学会研究報告. 自然言語処理研究会報告, Vol.2005, No.94, pp.73-78 (2005)

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8回)
SIG-AM-08-03

ブラウジング指向新聞アーカイブシステム KENBUN のための ズーミングインタフェース

A Zooming User Interface for the Browsing-based Newspaper Archive System KENBUN

後藤幹登　吉田翔一　中上裕基　萬卓哉　中島誠

Mikito Goto, Shoichi Yoshida, Yuki Nakagami, Takuya Yorozu, and Makoto Nakashima

大分大学工学部知能情報システム工学科
Department of Computer Science and Intelligent Systems, Oita University

Abstract: A zooming user interface for browsing-based archive system, KENBUN, is introduced to facilitate access the newspaper articles which are scanned from microfilm into digital images. The KENBUN provides browsing capability for the articles on the basis of the spatiotemporal knowledge of newspaper readers, where the digital images are linearly arranged according to their publish time on a two-dimensional space. We propose a zooming user interface for supporting both focused and contextual views of the space by providing the continuous and discrete interaction modes. The continuous and discrete interaction modes respectively allow the users move continuously on the space and interactively select certain times. The experiment revealed that using these modes is necessitated for browsing a large number of articles in finding not only a desired image but also text information.

1 はじめに

図書館が所蔵する新聞は，過去の出来事や歴史的背景を知る上で貴重な情報源の１つである[1]．望む記事への容易なアクセスが可能な新聞アーカイブシステムの必要性が注目されるにつれ[1,2]，図書館にマイクロフィルムで保存された紙面をスキャンニングにより画像化して，PC 上での閲覧を可能にする多くのプロジェクトが進められてきている[3, 4]．長期間に渡って発刊された大量の紙面画像がアーカイブされるが，アーカイブシステムが有効に働くには，紙面の特性を考慮した記事への柔軟なアクセス法の実現が重要となる．

望む記事へのアクセス法は，図書館内でユーザが開架書棚の間を歩き回り望む書物を探し出すようなブラウジングと，Web 上でユーザがキーワード等の質問用語を入力して，望むウェブページを探し出すような検索との 2 つに大別できる[5]．検索では，索引を利用することで，探索対象が大規模になったとしても質問に迅速に応じられる利点がある．しかしながら，質問用語がうまく与えられなければ，検索効率が低下してしまう．一方，ブラウジングでは，あらかじめ質問用語を用意しておかなくとも，探索対象を眺め回すだけで望みのものを見つけられる．予期せぬ発見があるという利点もある[5,6,7]．しかしながら，探索対象の規模が大きくなった場合への対処法を準備しておく必要がある．

これまで，新聞の発刊形態や記事の配置形態は時代が変わっても同じであるという事実を考慮に入れ，ブラウジングを主体として記事を探し出せる新聞アーカイブシステム KENBUN を構築してきた．新聞は時間の流れに沿ってほぼ毎日発刊されるが，人間は，時間を年，月，日といった単位にまとめこれらを順に並べて捉えている[8]．また，政治や社会といった同じ主題の記事は 1 枚のあるいは連続した紙面に，そして 1 つの記事は紙面ごとの段組み構成に沿ってまとめて配置されている．KENBUN では，このような時空間的知識を紙面や記事の扱いに関する背景知識とみなし，同じ発刊年月日の全紙面の画像を束ね，発刊年月日に従って構造化された 2 次元空間内に線形に配置している．紙面ごとの各記事は段組みに沿って紙面を矩形に分割して得た領域の連続部分に配置してある．

本論文では，上記のように配置された紙面画像のブラウジングを提供するためのズーミングインタフェースの設計について述べる．記事のブラウジングに際して，連続した時間の中で情報を探す場合もあれば，断続的にまとめた時間の中で情報を探す場合もある．ここでは，時系列順に沿った移動および紙面画像の拡大・縮小の連続性を有し，年月日のまと

まりごとの断続的なアクセスを可能にする２つのモードを有したズーミングインタフェースとする．

以下，新聞アーカイブシステムにおける紙面画像の配置と，ズーミングインタフェースについて，従来のシステムを概観し，次に提案するズーミングインタフェースの基本設計について述べる．さらにプロトタイプシステムの実現方法を述べ，有効性について検証する．

2 新聞アーカイブシステム

これまでに構築された新聞アーカイブシステムは，Google Newspaper Archive[9], 大英図書館[3]やオランダ王立図書館 [10]の新聞アーカイブシステムに代表されるように，紙面画像からのテキスト抽出を行い，検索による記事探索を中心として構築されている．テキスト抽出には，OCR（光学式文字認識）を用いるが，古い紙面であるほど抽出精度が低く，人手による修正が必要となる．ブラウジングに関しては，年，新聞社や地域などの指定により，紙面を限定して表示し，日を単位とした表示で，年月日の順序関係を一覧することができない．

一般に図書館では, 書籍をその扱う主題に注目し，十進分類表上で主題に割り当てられた数字の順に書棚に配置している．このような配置の利点は，近くの数字を与えられた書籍同士が関連性の高い主題を扱っていると言えることで，数字順に書籍を配置すれば，ブラウジングに適した配置となる．一方で，十進分類表における主題の並びについての理解がないと，図書館でのブラウジングを効果的に行うことは難しい．新聞は，様々な主題の記事を扱うことから，書籍と同様に配置することはできない．しかしながら，紙面をその発刊年月日順に配置すると，記事のブラウジングに，人間に共通の時間に流れに関する知識を利用することができる．主題ごとに関連した記事が物理的に隣接した空間に配置され，記事の並びは日々の読者であるユーザにとって馴染み深い．配置の理解に必要な時空間的知識をユーザがすでに有していると言える．

ズーミングインタフェース（Zooming User Interface: ZUI）[11]は，情報の配置の仕方が容易に理解できる場合の情報アクセス方法として効果を発揮する．大量の情報（画像，ウェブページやファイルなど）を俯瞰的に配置した空間内で，拡大あるいは縮小した空間を遷移しながら，詳細度の異なる情報をユーザがブラウジングできるシステムとして実現される．ズーミングの方法には，情報の特徴によって，拡大した空間と縮小した空間を連続的あるいは断続的に切り替える方法がある．

連続的なズーミングでは，代表的な Google Earth[12]のような地図を対象とするアプリケーションのみでなく，時間の流れに沿って情報を配置した年表をズーミングする ChronoZoom[13]がある．ChronoZoom では，ビックバンから現在までの 1 次元の時間の流れに沿った，宇宙の年表を表示し，ズーミングにより年代ごとのトピックを説明する情報を拡大する．画像の細部を拡大するように時間をズーミングすることで，年表のブラウジングを助ける仕組みとなっている．画像の詳細度に応じてテキストの詳細度も変化するため，テキスト自体は電子化されている必要がある．

断続的なズーミングは，ウェブディレクトリやファイル構造など，大量の階層的な情報を視覚化するシステムで用いられる. 代表的な例である, Zoomable Treemaps[14]では，各階層の 1 部分（ノード）を選択することでそのノードの詳細な情報を表示する空間へ遷移する. 隣のノードへの移動も可能であるが，ターゲットとするノードの選択を必要とし，同階層の近隣ノードを円滑に見渡すのに難がある．

紙面画像では，中に記されたテキストが情報源としての役割を持つだけでなく，段組みや，文字量，広告などが，画像としての情報源ともなり得る．テキストと画像双方の特徴を有する紙面画像のブラウジングのため，以下，上記のインタフェースの利点を取り入れたズーミングインタフェースを設計する．

3 ズーミングインタフェース

ここでは，時空間的知識によって捉えやすい，紙面画像の配置と，ブラウジングのためのズーミングインタフェースの基本設計について述べる．

3.1 KENBUN の紙面画像配置

図 1 に KENBUN における紙面画像の配置の仕方を示す．紙面画像は，時間の流れの認知しやすさを考慮すれば，日ごとの紙面を発刊年月日順に 1 次元に配置することが考えられる．しかしながら，図書館で書棚を配置できる場所が有限であるのと同様，ディスプレイ上の描画領域も有限である．図書館では，限られた場所で主題の並びを崩さず，ユーザの動線が連続するように書籍の配置が工夫されている．KENBUN では，1 カ月分のカレンダー様の表示を単位とし，2 次元空間上に左上から右下に時間の順に折り返しながら配置する. 日の部分には, 新聞同様，第 1 面の画像を一番上にして束ねて配置する．年を単位とするよりも折り返しに際して配置への自由度が高く, 日ごとに表示するよりも一覧性が高くなる．全ての月を一瞥できるように表示することで，どの

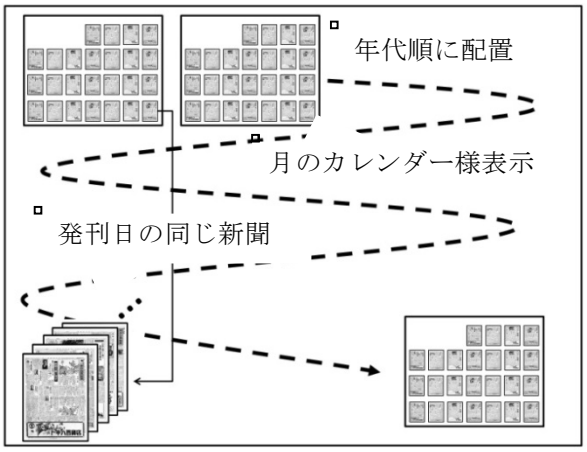


図 1　紙面画像の配置の方法

月に紙面画像があるか容易に判定できるようにしてある．

現在，KENBUN のプロトタイプは，大分県立図書館が収蔵する，昭和 38 年(1963 年）以前に大分県内および周辺地域で発刊された地方新聞のマイクロフィルムをスキャンニングして得た 156,686（朝刊：125,055，夕刊：31,005，号外：427，附録：199）ページ分の紙面画像を扱う．新聞社は合計で 25 社である．図 2 に紙面画像の実際の配置画面を示す．プロトタイプは，2012 年より大分県立図書館で稼働中である．

3.2 基本設計

KENBUN のために開発中のズーミングインタフェースの設計について述べる．大きく２つの部分よりなる．１つは，時系列順に沿った“**連続的**”ズーミングで，もう１つは，時間をまとめて捉える場合に備えた“**断続的**”ズーミングである．

連続的ズーミングの目的は，画像としての情報源の表示から，テキストとしての情報源の表示へと，詳細度だけでなく性質のことなる情報源の遷移を容易にすることにある．これは，図書館において，書棚にユーザが近づき，左右・上下に書棚を見渡す動作を基本とする．

KENBUN での紙面画像の配置は，2 次元空間上で

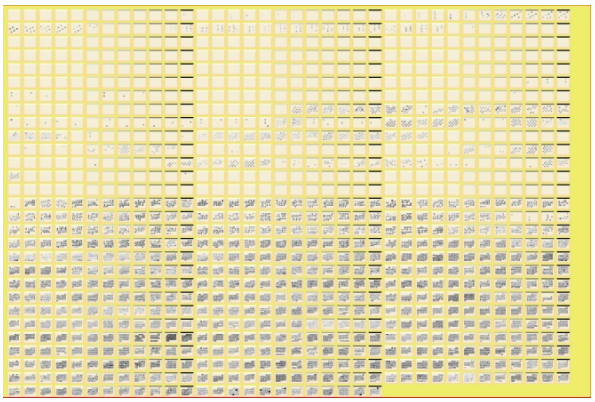


図 2　KENBUN の初期画面

はあっても，時系列のみの 1 次元配置で，図書館の書棚の配置と同様である．日常的な動作との連想を残せるように，連続的ズーミングの操作には，マウスホイールを利用して 1 次元的な動作で拡大と縮小ができるようにする．移動についても，マウスボタン押下後のドラッグにより直接的に行えるようにする．図 3 に，連続的ズーミングの過程を示す．マウスポインタの位置を中心として，ホイールの前後の操作量に応じて拡大率の異なる空間へと遷移させる．このとき，ユーザによる注目点の位置把握が混乱するのを避けるため，空間遷移は断続的ではなく連続的に行い，紙面画像のアスペクト比も変更しない．

断続的ズーミングでは，年代を想起しながら紙面画像を探索するユーザの支援のため，時間的なまとまりの階層ごとに，表示空間を遷移させて拡大する．時間的に近い空間も同時に拡大する．

図 4 に断続的ズーミングの過程を示す．マウスクリックに応じて，対応する年，その中の月，およびその中の日の拡大と進む．階層の遷移時に，ユーザの位置把握が混乱することを避けるためには，多くの研究で指摘されているように[15]，階層の遷移過程をアニメーションによって表示する．図のように，近隣にある月や日も同じ拡大率で表示することで，特定の時間のみでない情報を提示してブラウジングにおける新たな発見の利点を失わないようにする．

図 3　連続的ズーミングの過程

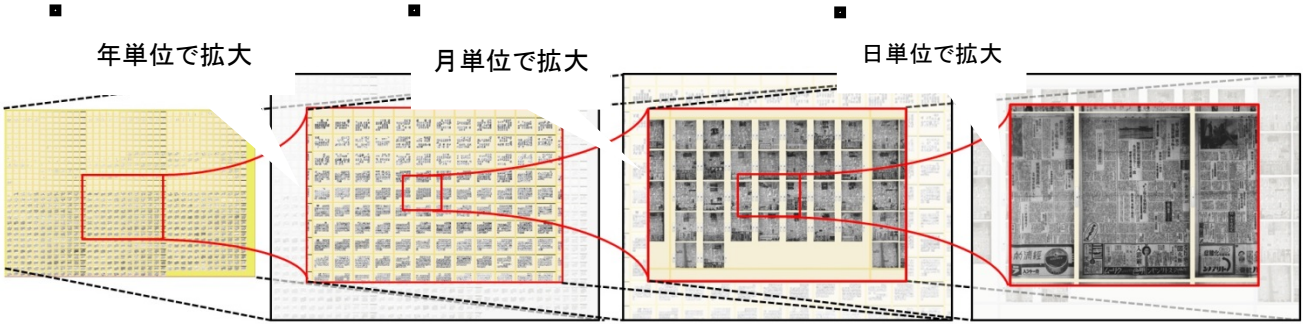


図 4　断続的ズーミングの過程

4 実現方法

前章で述べた 2 つのズーミングを可能にする機構の実現方法について述べる．

4.1 連続的ズーミング機構

連続的ズーミングでは，マウスポインタの位置を中心に，マウスホイールの操作量によって拡大した空間へと遷移する．また，2 次元空間上での上下左右の移動はマウスドラッグ操作によって行えるようにする．

ユーザの閲読に耐える高解像度の紙面画像を，動的に縮小して一覧表示するには，紙面画像の数が増えるほど計算コストが高くなり，動作に遅延を生じる．滑らかにズーミングを行うために，段階的に解像度の異なる紙面画像をあらかじめ用意しておき，マウスホイールの操作量に応じて，拡大する空間に用いるようにする．具体的には，図 5 に示すように「初期画面の段階」,「年単位の拡大時」,「月単位の拡大時」,「日単位の拡大時」,「記事閲覧時の段階」を合わせた計 5 段階分を用意し，拡大率に応じて表示する紙面画像を切り替えていく（図中の数字は，段階ごとの紙面画像の解像度である）．ただし，画像の切り替えがユーザにわからないように，例えば A 段階から B 段階への拡大時の切り替えでは，A 段階の縮小率と B 段階の拡大率を調整して，見かけ上の紙面画像の大きさを一致させる．その際，A 段階の表示されている紙面画像のみに関して，B 段階の画像を読み込む事で負荷軽減を行っている．

画像の読み込み回数を減らすために，初期画面を表示する際には，図 6 で示すような新聞をカレンダー上に配置し繋げた 1 枚のカレンダー画像を用意し，さらにそれらのカレンダー画像を繋げた一枚の画像を用意する．これにより，起動時の画像の読み込みは 1 度で済む．

4.2 断続的ズーミング機構

断続的ズーミングでは，マウスクリックによる年月日の選択により，それらを見渡せる解像度の紙面画像を用いた空間へと遷移させる．結果的に月のカレンダー様の表示および, 日の拡大となる. ここで，連続的ズーミング機構におけるドラッグ操作との競合を避けるため，ダブルクリック操作によってズーミングを行う．また，拡大の際に，周りの年，月，日の紙面画像も表示して，同一階層間での移動も可能とする．ひとつ前の階層に戻るには（つまり，縮小した空間への遷移には), 紙面画像の間をクリックすればよく，操作時のマウスポインタの移動距離が少なくて済む.

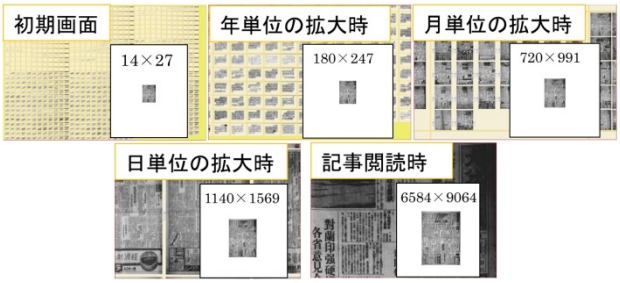


図 5　段階ごとの新聞紙面の解像度

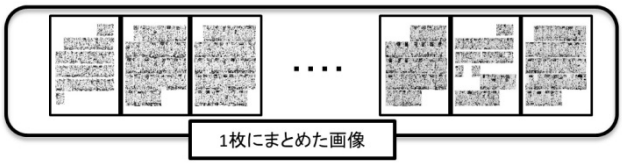


図 6　起動時に読み込む画像

拡大あるいは縮小した空間への遷移時のアニメーションは，従来の研究 [16]を参考に 1 秒以内で終了させるようにして，遅延とのトレードオフに配慮する．また，画面の上辺に年代を表すルーラーを表示し，マウスポインタ下の新聞紙面の発刊日が全体の中のどの位置にあるかを表示する.

「日単位の拡大時」では，図 7 に示すようなページめくりボタンを紙面画像の左右に表示する．多数のページを有する情報源では，書籍のように，日頃慣れた動作であるページめくり動作が読み易いとさ

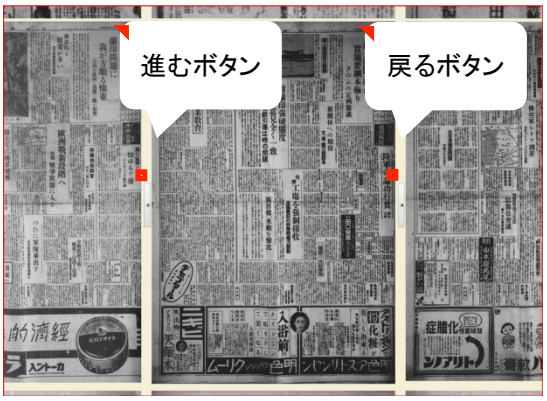


図 7　ページめくりボタン

れる．ボタンの位置と対応する機能は，実際の新聞を読む動作を考慮しており，左のボタンを押すことでページが進み，右のボタンを押すことでページが戻るようにしてある．

5 評価実験

提案したズーミングインタフェースのプロトタイプを Web アプリケーションとして実装し，紙面画像のブラウジング時の利用のされ方を，被験者実験によって検証した．Google Chrome をブラウザとして用いた．2 つのズーミングの使われ方をみるため，表示する紙面は 2 年分とした．「年」「月」「日」の階層構造を利用する最小の量である．

11 名の大学生を被験者として，連続的，断続的ズーミングの操作方法を練習してもらったうえで，以下の 3 つタスクを行ってもらった．ズーミングの過程のログを記録し，連続的ズーミングではマウスホイール，断続的ズーミングではマウスクリックについてその操作回数を計測した．

タスク 1：特定の紙面画像の探索
タスク 2：「天皇」に関する見出しの探索
タスク 3：気になった見出しを 5 つ抽出

表 1　ズーミング操作の比率

ズーミング	タスク 1	タスク 2	タスク 3
連続的	0.56	0.60	0.67
断続的	0.44	0.40	0.33

3 つのタスク完了後，被験者には以下の項目に答えてもらった．Q1 と Q2 の回答は，5 段階のリッカート尺度（5. 非常に同意できる，4. 同意できる，3. どちらともいえない，2. 同意できない，1. 全く同意できない）とした．

Q1 ズーミングの操作方法は分かり易かった
Q2 アニメーションの速さは適切だった
Q3 どちらのズーミングが使いやすかったか

表 1 に，全体の操作数に対する各ズーミングの操作回数の比率の平均を示す．タスク間で，比率に統計的有意な差はなかった（1 次元配置分散分析，有意水準 0.05．タスク間で比率の分散にも有意差はなかった）．また，全般的に断続的ズーミングの操作比率が少ないが有意な差はない．いずれのズーミングだけが用いられるわけではなく，ユーザの嗜好に応じてどちらも利用されることが分かる．

図 10 に，操作時間が最も長かったタスク 3 で，典型的な被験者のズーミングの操作過程を示す．縦軸には，上から，断続的ズーミングで初期画面に戻るクリック（全体クリック），年，月および日単位の画面を表示するためのマウスクリック，ドラッグ操作，連続的ズーミングでの拡大と縮小を行うマウスホイールの操作を示す．横軸は，全操作時間で正規化した経過時間である．断続的ズーミングを行った後，ドラッグ操作による移動をする被験者はほとんど無く，連続的ズーミングの後行っていた．紙面画像を確実に見るために，被験者が望む拡大率を連続的ズーミングで調整したためである．断続的ズーミングでユーザの望む拡大率にするには，異なる解像度の

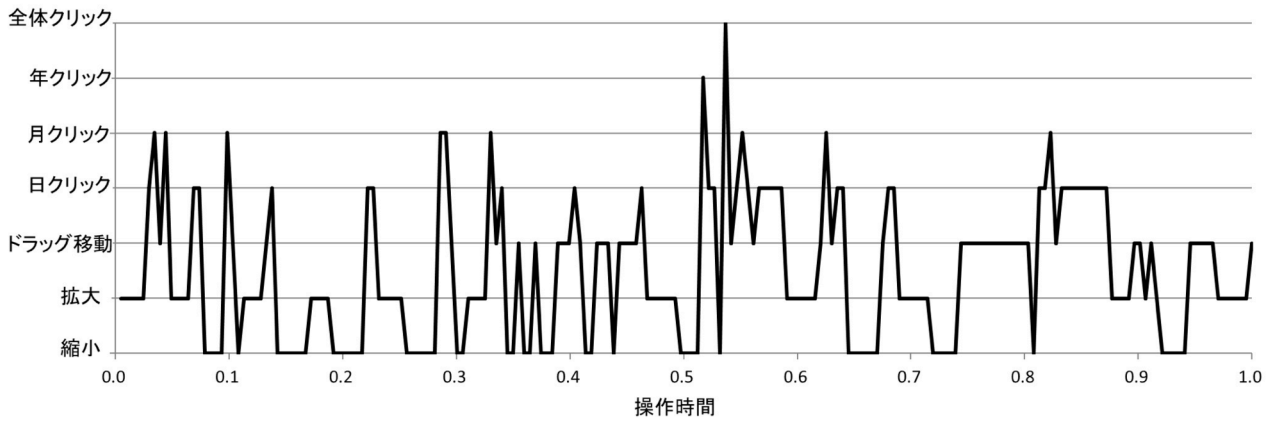


図 10　ズーミング操作過程

表 2 アンケート結果（人）

	回答				
	1	2	3	4	5
Q1（連続的）	0	0	1	3	7
Q1（断続的）	0	3	4	3	1
Q2	0	4	3	2	1
	連続的		どちらでもない		断続的
Q3	7		2		2

画像を多く用意する必要があるが，多くなればそれだけ，読み込みのコストが高くなる．連続的ズーミングは効果的なブラウジングに必須であることが分かる．

表 2 にアンケートの結果を示す．Q1 で，連続的ズーミングが使い易いとした意見が有意に高かった（回答 4,5 を肯定，1,2 を否定意見とした 2 項検定，有意水準 0.05）．断続的ズーミングでは，Q2 の結果から，アニメーションの速度に問題があるとわかる．アニメーション時に文字と画像の両方を目で追うために疲労を感じるというユーザ意見もあり，ユーザの負担を減らせるような改善が必要である．しかしながら，Q3 で，ユーザの好みに有意な差はなく，双方のズーミングの用意は必要であることが分かる．

6 おわりに

ブラウジング指向新聞アーカイブシステム KENBUN のためのズーミングインタフェースの提案と検証を行った．今後，連続的ズーミングにおける画像切り替え速度の向上と断続的ズーミングのアニメーションの動作についての検証を重ね，ブラウジングをより円滑に行えるシステムとして実装する．また，KENBUN への導入とともに，図書館におけるユーザテストを通じた評価を行っていく．

7 参考文献

[1] 佐々木美穂: 英国とオランダの国立図書館にみる新聞資料デジタル化プロジェクト, カレントアウェアネス, no. 309, pp. 2-5, (2011).

[2] Balk, H., and Conteh, A.: IMPACT: Centre of competence in text digitization, Proc. 2011 Workshop on Historical Document Imaging and Processing (HIP'11), pp. 155-160, (2011).

[3] British Library, British Newspaper Archive. http://www.britishnewspaperarchive.co.uk/

[4] Ishihara, T., Itoko, T., Sato, D., Tzadok, A., and Takagi, H.: Transforming Japanese archives into accessible digital books, Proc. JCDL'12, pp. 91-100, (2012).

[5] White, R. W., Kules, B., Drucker, S. M., and Schraefel, M. C.: Supporting exploratory search: Introduction, Communications of the ACM, vol. 49, no. 4, pp.36-39, (2006).

[6] Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., Yee, K.-P.: Finding the flow in Web site search, Communications of the ACM, vol. 45, no. 9, pp. 42-49, (2002).

[7] 鈴木明,他:多様な情報資料へのアクセスに対応した「ブラウジング」環境…大学図書館の新しい環境整備手法に関する研究, 神戸芸術工科大学紀要 2009「芸術工学」, (2009).

[8] 中島義道,「時間」を哲学する，講談社現代新書，(1996).

[9] Google Newspaper Archive. http://news.google.com/newspapers

[10] National library of the Netherlands, Databank of Digital Daily newspapers. http://www.kb.nl/hrd/digi/ddd/index-en.html

[11] Bederson, B. B.: Pad++: Advances in multiscale interfaces, Conference Companion on Human Factors in Computing Systems, pp. 315-316, (1994).

[12] Google Earth. https://www.google.co.jp/intl/ja/earth/

[13] Walter, R. L., Berezin, S., and Teredesai, A.: ChronoZoom: Travel through time for education, exploration, and information technology research, Proc. RIIT'13, pp. 31-36, (2013).

[14] Blanch, R., and Lecolinet, E.: Browsing zoomable treemaps: Structure-aware multi-scale navigation techniques, IEEE Transactions on Visualization and Computer Graphics, vol. 13, no. 6, pp.1248-1253, (2007).

[15] Bederson, B.: The promise of zoomable user interfaces, Behaviour & Information Technology, vol. 30, no. 6, pp. 853-866, (2011).

[16] Cockburn, A., Karlson, A., and Bederson, B. B.: A review of overview+detail, zooming, and focus+context interfaces, ACM Computer Surveys, vol. 41, no. 1, pp. 1-31, (2008).

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8回)
SIG-AM-08-04

オンラインニュースを対象とした モニタリングシステムの提案

Proposal of Monitoring System for Online News

沼野 航希* 高間 康史

Koki Numano, Yasufumi Takama

首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

Abstract:本稿では，オンラインニュースの定期的なモニタリングを支援する情報可視化システムを提案する．オンラインニュースは主要な情報源の一つとなっているが，新着記事が絶え間なく到着し，常時モニタリングすることは困難である．興味ある話題を見逃すことなく効率的にモニタリングするために，提案システムでは以前関心を抱いた話題の続報提示，モニタリングするタイミングを判断する手がかりの提示によりモニタリングを支援する．

1. はじめに

本稿は，オンラインニュースの定期的なモニタリングを支援する情報可視化システムを提案する．近年，オンラインニュースは Web 上で，主要な情報源の一つとなっている．ニュースサイトは多数存在し，2014 年 10 月 2 日の新着記事数は，「朝日新聞デジタル[1]」が 129 件，「日本経済新聞[2]」が 270 件，「毎日新聞[3]」が 112 件であった．このように，ニュースサイト一つあたりの新着記事件数は 100 件を超えるが，複数のニュースサイトを閲覧することが一般的であるため，一人のユーザが一日に受け取る新着記事は数百件になることも珍しくない．ユーザは日常生活において，これらの記事を継続的に全てモニタリングすることは困難であるため，モニタリングしていない間の情報の見逃しが発生することが問題としてあげられる．そのため，ユーザが関心を抱いている話題を効率的にモニタリングできるようにすることが重要と考える．

効率的なニュース閲覧を支援するサービスとして，ニュースキュレーションサービスが急速に普及しつつある．代表的なニュースキュレーションサービスの一つである「グノシー[4]」は，独自のアルゴリズムでユーザの興味に合った最新ニュースを提示する他，時間指定によるプッシュ通知お知らせなどの機能があるが，適切なタイミングでニュースを確認できているか否かは考慮されていない．

本稿で提案するシステムは，話題検出・追跡技術を用いてユーザが関心を抱いている話題を可視化して提示する．また，ユーザがモニタリングするタイミングを判断する手がかりも可視化して提示する．本稿では，構築したプロトタイプシステムについて述べるとともに，ユーザに利用してもらった予備実験の結果について報告する．

2. 関連研究

2.1 ニュースキュレーションサービス

ニュースキュレーションサービスとは，Web 上のニュースを収集，分類を行いユーザに提供するサービスのことである．スマートフォンの普及に伴い，急速に利用者が増大している．代表的なキュレーションサービスに前述のグノシー[4]の他，SmartNews[5]などが挙げられる．

SmartNews は，エンタメ，スポーツ，グルメなど 11 のジャンルの中から，読みたい話題を自由に選択

[1] http://www.asahi.com/
[2] http://www.nikkei.com/
[3] http://mainichi.jp/

*連絡先:首都大学東京大学院システムデザイン研究科
〒191-0065 東京都日野市旭が丘 6-6
E-mail:ytakama@sd.tmu.ac.jp

[4] http://gunosy.com/
[5] http://www.smartnews.com/ja/

し，並び替えることができる．また，Twitter でツイートされた Web ページをリアルタイムで解析し，話題になっている記事を配信する機能も備えている．

2.2 話題検出・話題追跡

時系列に到着する一連のニュースなどから新規に出現した話題を抽出することを話題検出，既出話題の続報を検出することを話題追跡と呼ぶ．テキストデータを対象とした話題検出・追跡の手法は様々に提案されている[2][3][4][5][6][7][8][9][10]．ニュース記事のようなテキストデータを対象とする場合，記事間または，記事と記事クラスタ間の類似度を求めることにより話題の抽出をする手法が一般的である．話題検出・追跡処理の一般的な流れを以下に示す．

1. 特徴量の計算…各記事から特徴ベクトルの生成
2. 文書クラスタリング…話題に対応したクラスタの生成

ステップ 1 では，クラスタリングを行う前処理として特徴量の計算を行う．記事及びクラスタの表現は，ベクトル空間モデルがよく用いられる[2][3][4][5]．ベクトル空間モデルでは，単語の重みは tfidf で求めることが多いが，上嶋ら[5]は，idf 値を更新することは，一度決定した過去のクラスタリング基準が変わってしまう場合があるという理由から tf 値のみを用いて逐次クラスタリングを行っている．菊池ら[2]は，過去の文書から事前に求めた idf 値を用いている．

ステップ 2 で行う文書クラスタリングの手法も様々に提案されている．クラスタの重心ベクトルと文書ベクトルの類似度を余弦尺度を用いて計算し，逐次クラスタリングにより話題クラスタを抽出する手法[2][3][4][5][6]，共起語集合が話題を形成するとの考えに基づき，共起語集合間の類似度 JS divergence を用いて計算し話題を抽出する手法[7]などが提案されている．JS divergence とは，2 つの分布の相違度を測る尺度である KL divergence を対称化したものである．0 から 1 までの値をとり，値が大きいほど 2 つの分布は異なっている．

芹澤ら[4]は，コサイン類似度を用いて各トピック間の類似度を求め，連続する 2 日間の類似度が閾値以上ならばトピック間に関連付けを行うことでトピックを追跡する．

2.3 可視化表示システム

文書クラスタリングによって生成された話題クラスタをわかりやすくユーザに提示するために，時系列ごとに話題の遷移を示すインタフェースも様々に提案されている[3][6]．

森ら[3]は，2 次元平面上の横軸に時間軸を，話題クラスタを縦軸に配置して話題遷移を可視化する手法を提案している．話題の分岐，収束の両方を確認することができ，前後関係や話題の追跡が容易になるとしている．

長谷川ら[6]が提案する T-Scroll は，時系列文書を対象とするクラスタリングシステムが定期的に生成するクラスタリング結果をもとに，クラスタ間の関連を巻物状に可視化する．楕円でクラスタを示し，その中には，そのクラスタを最も適切に表すようなキーワードを選んで表示する．また，楕円のマウスオーバー時に，クラスタに含まれる文書一覧を表示する機能も備えている．楕円どうしを繋ぐことにより，話題の時系列変化を把握し，クラスタの内容を容易に確認できることがこのシステムの特徴である．

3. オンラインニュースを対象としたモニタリングシステム

本稿では，オンラインニュースの定期的なモニタリングを支援する情報可視化システムを提案する．具体的には，前回モニタリング以降に到着したオンラインニュースについて，新規に発生した話題に関する記事，前回関心を持った話題の続報記事の発見を支援する．提案システムは，オンラインニュースの収集, 文書クラスタリングによる話題検出･追跡, インタフェースによる提示から構成される．図 1 にシステム構成図を示す．以下では構成要素それぞれについて説明する．

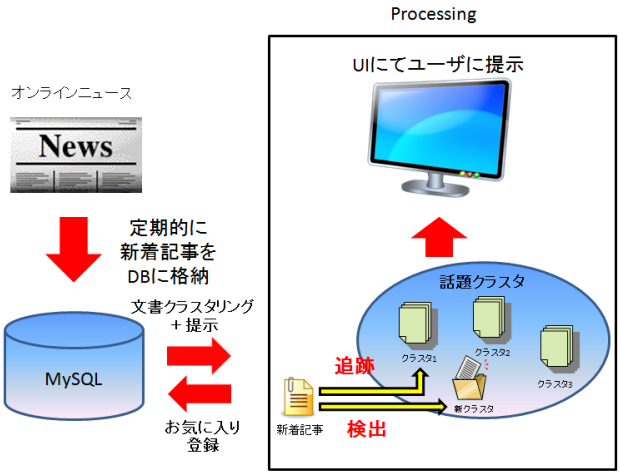


図 1. システム構成図

3.1 オンラインニュースの収集

モニタリングシステム構築にあたり，オンラインニュースを Web 上から収集する．オンラインニュースに含まれる情報は title，date（配信日時），text（記事本文）であり，これらの情報を得るために，Ruby の rubygems ライブラリである Mechanize を使用する．4 節で述べる実験では，朝日新聞デジタル[6]の新着記事を 2014 年 6 月 1 日～6 月 30 日の期間取得して，記事ごとにデータベースに格納したものを用いている．表 1～表 3 にデータベースの構成について示す．newstable（表 1）は，記事内容と記事の配信日時を格納するためのテーブルである．apclustertable（表 2）は，話題クラスタ毎の記事番号を格納するためのテーブルである．favoritefeednotable（表 3）は，ユーザが記事および話題クラスタをお気に入り登録した際に，記事番号，配信日時，お気に入り登録された回数を格納するためのテーブルである．インタフェースで提示する際に，関心のある話題クラスタに関して配信日時の情報を必要とするため，favoritefeednotable にも date を格納する．

表 1. newstable

カラム	内容
id	記事番号（1 から順に auto_increment）
title	記事のタイトル
date	記事の配信日時
text	記事本文

表 2. apclustertable

カラム	内容
clusterno	クラスタナンバー （1 から順に auto_increment）
feedno	記事番号（newstable の id）

表 3. favoritefeednotable

カラム	内容
id	記事番号（newstable の id）
date	記事の配信日（newstable の date）
count	お気に入り登録した回数

3.2 文書クラスタリング

文書クラスタリングには Affinity Propagation[1]アルゴリズムを用いる．Affinity Propagation アルゴリズムは，予めクラスタ数を決めておく必要がなく，クラスタリング結果が初期値に依存しないという特徴を持っている．本稿で対象とするオンラインニュースは時系列的に発生するため，予めクラスタ数を決めることができないことから，クラスタリング手法に Affinity Propagation アルゴリズムを用いた．Affinity Propagation アルゴリズムは，全要素間の関係性を similarity として設定し，availability と responsibility という 2 種類のメッセージを交換し合うことで，exemplar（クラスタの中心）を決定し，クラスタを生成する手法である．responsibility($r(i,j)$), availability($a(i,j)$)は共にデータポイントが j が i の exemplar としてふさわしい度合いを表すが，前者は i が j を選ぶ度合いであり，i から j へ送られるのに対し，後者は j が i にとってふさわしい度合いであり，j から i に送信される．Affinity Propagation アルゴリズムのフローチャートを図 2 に示す．

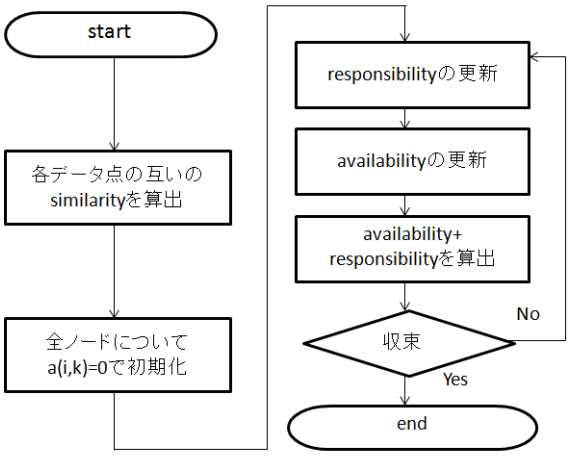


図 2. Affinity Propagation アルゴリズムのフローチャート

本稿では，similarity の算出に cos 類似度を用いる．similarity 算出の手順を示す．

1. newstable の title と text を形態素解析し，名詞と未知語を抽出
2. 抽出した単語を特徴とし，tfidf 値を重みとして，各新聞記事の特徴ベクトルを生成
3. 全記事対の cos 類似度を計算

ステップ 3 の結果に基づき，newstable に格納されている記事数を N として $N \times N$ の類似度行列を生成し，Affinity Propagation アルゴリズムに入力する．

Affinity Propagation アルゴリズムの収束条件は，クラスタ割り当てが直前の結果と変化がない場合，

[6] http://www.asahi.com/

または計算の反復回数が最大値を超える場合である．今回は予備実験の結果に基づき，反復回数を 50 回としてクラスタリングを行った．

availability および responsibility は以下の式で算出される[11].

$$r(i,j) = (1-\lambda) * \rho(i,j) + \lambda * r(i,j) \quad (1)$$
$$a(i,j) = (1-\lambda) * \alpha(i,j) + \lambda * a(i,j) \quad (2)$$

ここで，λは Damping Factor と呼ばれる，反復計算の中で availability および responsibility が振動するのを防ぐための係数である．今回は，予備実験の結果に基づきλ=0.9 で反復計算を行った．また，ρ(i,j)とα(i,j)は以下の式から計算する．

$$\rho(i,j) = \begin{cases} s(i,j) - \max_{k \neq j}\{a(i,k) + s(i,k)\} & (i \neq j) \\ s(i,j) - \max_{k \neq j}\{s(i,k)\} & (i = j) \end{cases} \quad (3)$$

$$\alpha(i,j) = \begin{cases} \min\{0, r(j,j) + \sum_{k \neq i,j} \max\{0, r(k,j)\}\} & (i \neq j) \\ \sum_{k \neq i} \max\{0, r(k,j)\} & (i = j) \end{cases} \quad (4)$$

3.3 インタフェース

3.2 節に示した文書クラスタリング結果に基づき，話題クラスタをユーザに提示するためのインタフェースを提案する．開発には Processing を用いた．MySQL Server から新聞記事及び話題クラスタに関する情報を取得し，Processing にて提示する．提案するインタフェースには，続報記事数提示モード，リストモードの二つのモードがある．各モードのスクリーンショットを図 3，4 にそれぞれ示す．

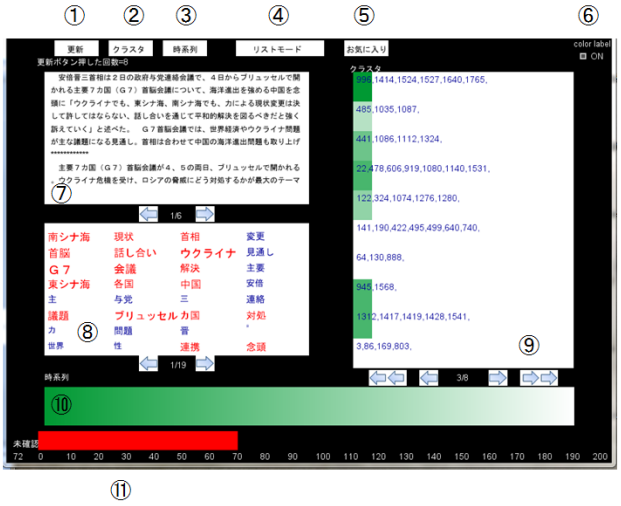


図 3. インタフェース（リストモード）

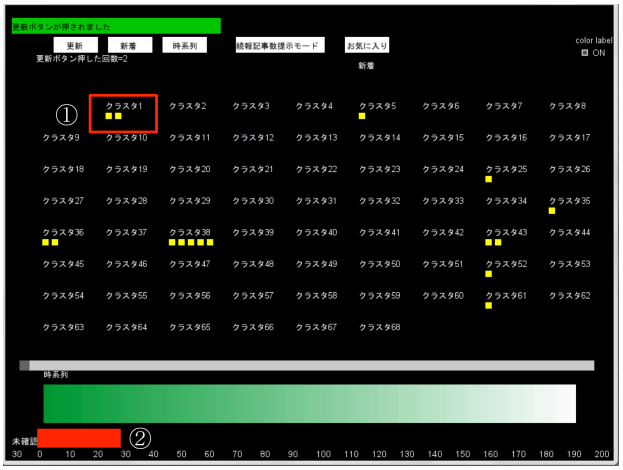


図 4. インタフェース（続報記事数モード）

ユーザはこのインタフェースを使用して，新着記事及び話題クラスタ毎の記事内容の確認や，関心のある話題クラスタの続報記事数の確認を行う．

リストモード（図 3）では，新着記事及び話題クラスタ毎の記事内容の確認を行い，続報記事数提示モード（図 4）では，関心のある話題クラスタの続報記事数の確認を行う．それぞれのモードについて，以下に機能の説明を示す．

リストモード（図 3）では，新着記事の見出し，または話題クラスタ毎に記事番号を確認できる（図 3 の⑨）．新着記事⇔話題クラスタの切替は，新着⇔クラスタ切替ボタン（図 3 の②）で行う．リスト（図 3 の⑨）内のテキストをクリックすると，新着記事の場合は記事内容が表示され（図 3 の⑦），話題クラスタの場合は，クラスタ内の記事内容（図 3 の⑦）とクラスタ内の単語についてのタグクラウドが表示される（図 3 の⑧）．新着記事及び話題クラスタリストで関心を抱いたものがあれば，お気に入りボタン（図 3 の⑤）により，お気に入り登録を行う．

また，リスト内の着色の方法に関して，時系列⇔関心の有無切替ボタン（図 3 の③）により，時系列または関心の有無のどちらを基準に着色するのかを切り替えることができる．新着記事⇔話題クラスタ切替ボタンとの組み合わせにより，以下の 3 機能が利用可能である．ここで，新聞記事と時系列は同種の情報であるためその組み合わせは除外している．

機能 1. 新着+関心の有無
→新着（未確認）記事の中で，前回関心を持っていた記事の続報を強調
機能 2. クラスタ+時系列
→各クラスタの最新記事を比較：より最新の記事を含むクラスタを強調

機能 3. クラスタ+関心の有無
→前回お気に入り登録した記事が多く含まれるクラスタを強調

続報記事数提示モード（図 4）では，話題クラスタ毎に，ユーザがまだ内容を確認していない続報記事数を黄色い四角で提示する（図 4 の①）．なお，続報記事数が提示されるクラスタは，ユーザが前回お気に入り登録をした記事が含まれるクラスタのみである．また，ユーザが前回記事内容を確認してから到着した新着記事数を赤いバーとともに表示する（図 4 の②）．

リストモードと続報記事数提示モードの切替は，リストモード⇔続報記事数モード切替ボタン（図 3 の④）により行う．

4. 評価実験

4.1 実験概要

本実験では，20 代の工学系大学院生を対象に，提案したシステムのプロトタイプを使用してモニタリングを行ってもらった．本実験の検証目的は，短時間で関心のある記事を確認できること，記事内容を確認する必要があるか否かを判断できることの 2 つである．3.1 節で述べた通り，実験には，朝日新聞デジタルの新着記事[7]を 2014 年 6 月 1 日～6 月 30 日の期間取得して用いた．モニタリングは本務の合間に行われるとの想定に基づき，実験協力者には他の作業を適宜してもらいながら，提案システムを利用してもらった．以下に実験手順を示す．

1. 更新ボタンを 1 回押して新着記事あるいは新規クラスタから 5 個ずつ計 10 個お気に入りに登録
2. 実験開始時間から 5 時間の間に，ユーザの任意のタイミングで続報記事数を確認（回数は 13～15 回）
3. 記事内容の確認が必要か否かを判断
必要と判断した場合→4-(a)へ
必要でないと判断した場合→4-(b)へ
4-(a). 5 分の制限時間で記事内容を確認し，適宜お気に入り登録
4-(b). 他の作業の再開
5. Step2～Step4 をユーザが制限時間内に繰り返し試行

実験において，ステップ 3-(a)の記事内容確認は 5 回に限定した．これにより，新着記事があまりない状態で確認してしまうと，後の方で多数の記事を一度に確認しなくてはならない状況が発生することになる．従って，5 分間で確認できる程度の新着記事が到着した，適切なタイミングを実験協力者が判断可能かどうかが検証可能と考える．

[7] http://www.asahi.com/

4.2 実験結果

表 4 に, 各実験協力者の記事内容確認時間を示す. 表 5 に，関心のある話題に関して追跡ができた割合として，続報記事を含む話題クラスタを確認した割合（左セル），続報記事を確認した割合（右セル）を示す. 表 6 に, 前回の確認時から到着した記事数 (左セル)，続報記事数（右セル）を示す.

表 4. 記事内容確認時間

	A	B	C	D
1 回目	4 分 01 秒	2 分 42 秒	5 分 44 秒	5 分 52 秒
2 回目	3 分 43 秒	3 分 02 秒	9 分 59 秒	7 分 00 秒
3 回目	2 分 38 秒	3 分 25 秒	7 分 13 秒	7 分 45 秒
4 回目	1 分 53 秒	4 分 51 秒	7 分 16 分	8 分 32 秒
5 回目	1 分 41 秒	3 分 52 秒	5 分 49 秒	4 分 01 秒

表 5. 話題追跡できた記事数の割合
（左：話題クラスタ確認，右：続報記事確認）

	A		B		C		D	
1 回目	0.67	0.40	0	0.57	0.33	0.78	1.00	0.22
2 回目	1.00	0.50	0.71	0.10	1.00	0.92	1.00	0.79
3 回目	0.10	0.10	1.00	0.23	1.00	0.97	1.00	0.93
4 回目	0	0.08	1.00	0.32	1.00	1.00	0.93	1.00
5 回目	0.18	0.08	1.00	0.22	1.00	0.88	0.70	0.86

表 6. 未確認の新着記事到着件数
（左：到着記事数，右：続報記事数）

	A		B		C		D	
1 回目	60	5	80	7	80	9	111	18
2 回目	51	10	81	10	132	13	50	14
3 回目	160	30	72	13	101	33	72	14
4 回目	100	13	138	31	96	28	212	43
5 回目	74	13	74	27	118	60	40	14

実験結果より，実験協力者 A，B は全ての回に 5 分以内で記事内容を確認しているのに対し，C，D はほとんどが 5 分を超えていることがわかる．C，D は，到着した続報記事，続報を含む話題クラスタを高い割合で確認していることが，記事確認に多くの時間を要した原因であると考える．

適切なタイミングでモニタリングを行えているか

否かに関して，実験協力者毎の記事内容確認時間のばらつきについて考察する．1 回目はモニタリング開始のため 2 回目以降よりも時間がかかること，および 5 回目は実験終了の制約があることを考慮して，2〜4 回目のモニタリングにおける確認時間の最大値，最小値の差を見ると，実験協力者 A は 1 分 50 秒，B は 1 分 49 秒，C は 2 分 46 秒，D は 1 分 32 秒であった．実験協力者 C は確認時間が長いことを考慮すれば，実験協力者によらず確認時間のばらつきは大きくなく，適切なタイミングで確認が行えていると考える．また，実験協力者 A 以外は，1 回目を除き 70%以上関心のある話題クラスタあるいは続報記事を確認できていることがわかる．1 回目は，インタフェースに関して操作の要領を得ていないことが原因として考えられる．A は，関心を持った話題クラスタあるいは新着記事を効率よく見つけられなかった可能性がある他，それらの話題に対して次の確認時興味を失い，確認をしなかった可能性もあるため，今後調査の必要があると考える．

実験後，話題検出・追跡，提示するタイミング，システム全体のことに関して，アンケートを実施した．話題検出・追跡に関しては，ほとんどが続報を確認できたと回答しているが,「サッカー」という話題でも，自分の関心の無い国に関する話題も続報としてまとめられていたため，全ての続報に関心を持ったとは言えないという意見があった．データ収集時期がワールドカップ開催期間であったため，サッカーに関する話題として広い範囲で一つの話題クラスタが生成されてしまったことが原因であると考える．この問題を解決するには，より時系列を考慮したクラスタリング手法が必要と考える．提示するタイミングに関しては，ほとんどが適切なタイミングで記事内容を確認できたと回答している．システム全体に関しては,画面の遷移がわかりにくいことや,どの記事がどの記事の続報であるかわかりにくいなどの意見があった．話題追跡に関して，A 以外は高い割合でできていることから，インタフェースの完成度を上げることで，より短時間で記事内容を確認できることが期待できる．

5. おわりに

本稿では，オンラインニュースの定期的なモニタリングを支援する情報可視化システムを提案した．提案するシステムでは，ユーザが関心を抱く話題の追跡だけでなく，話題クラスタごとに続報記事数を提示することで，より適切なタイミングで記事内容を確認することを支援する．プロトタイプシステムを構築し，評価実験を行った結果，高い割合で関心のある話題クラスタについて追跡できることを示した．

今後は，時系列を考慮したクラスタリング手法に改善するとともに，インタフェースの操作性を向上し完成度を高めることで，より見やすいインタフェースを検討する予定である．

参考文献

[1]B. J. Frey and D. Dueck: Clustering by passing messages between data point, Science, Vol. 315, pp. 972-976, 2007.
[2]菊池匡晃，岡本昌之，山崎智弘：階層型クラスタリングを用いた時系列テキスト集合からの話題抽出, DBSJ Journal, Vol. 7, No. 1, pp. 86-90, 2008.
[3]森幹彦：ニュース記事の話題分岐を時系列で追跡可能な可視化法，情報処理学会第 71 回全国大会，6B-3，2009.
[4]芹澤翠，小林一郎：潜在トピックの類似度に基づくトピック追跡への取り組み, 第 25 回人工知能学会全国大会，3F3-2，2011.
[5]上嶋宏，三浦孝夫，塩谷勇：時系列ニュース記事集合に基づくニュース記事の順序付け, DEWS2004, 1-B-04，2004.
[6]長谷川幹根，石川佳治 : T-Scroll : 時系列文書のクラスタリングに基づくトレンド可視化システム，情報処理学会論文誌 : データベース，Vol. 48，No. SIG 20 （TOD 36），pp. 61-78，2007.
[7]森井洸明，アダムヤトフト，田中克己：ニュースアーカイブを用いた話題変化と原因語の発見，DEIM Forum 2012，D4-2，2012.
[8]橋本泰一，村上浩司，乾孝司，内海和夫，石川正道：文書クラスタリングによるトピック抽出および課題発見, 社会技術研究論文集, Vol. 5, pp. 216-226, 2008.
[9]高橋祐介，横本大輔，宇津呂武仁，吉岡真治，河田容英，神門典子，福原知宏，中川裕志，清田陽司：時系列トピックモデルにおけるバーストの固定，DEIM Forum 2012，F5-5，2012.
[10]平田紀史，大囿忠親，新谷虎松：ユーザの選好に基づくトピック分析システムの試作，JSAI2008, pp. 277-277，2008.
[11]藤原靖宏，入江豪，北原友恵 : Affinity Propagation のための高速化手法，DEIM Forum，C1-3，2012.

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8回)
SIG-AM-08-05

複数国の新聞からの多観点比較による分析 ～GDELT データを用いた分析～

Multifacet comparative analysis of newspaper articles from different conutries - Analysis based on Global Database of Events, Language and Tone (GDELT) -

吉岡真治 [1*]　神門典子 [2]
Masaharu Yoshioka[1]　Noriko Kando[2]

[1] 北海道大学大学院情報科学研究科
[1] Graduate School of Information Sciecne and Technology, Hokkaido University
[2] 国立情報学研究所
[2] National Institute of Informatics

Abstract: The News Site Contrast (NSContrast) system analyzes multiple news sites based on the concept of contrast set mining and it can extract terms that characterize different topics of interest for specific countries. In this study, we used the NSContrast system to analyze Global Database of Events, Language and Tone (GDELT) data by comparing news articles from different regions (e.g., USA, Asia, and the Middle East). We also present examples of analyses performed using this system.

1 はじめに

近年、世界中の報道機関が、各々の記事を配信するニュースサイトを構築し、記事を提供している。その結果、これまでは、簡単には読むことの出来なかった世界中のニュース記事を閲覧することができるようになってきた。我々は、これまでに、複数の複数の国のニュースサイト発信する記事を比較することにより、それぞれのサイトが代表する各国の興味の違いなどを分析する新聞比較システム NSContrast の開発を行ってきた [1, 2]。このようなシステムを構築する上での問題の一つは、大規模なニュースサイトから記事を集め、分析を行う必要があるという点にあった。

この様なニュースの分析をするというニーズに答えるための基盤データとして、Global Database of Events, Language and Tone (GDELT) [3] [1]が、2013 年から公開されている。このデータベースは、Google News に提供されている様々なニュースサイトのデータなどを用いて構築された非常に大規模なデータベースで、世界中で報道されている毎日の記事において、どの様なイベントが、どの様なトーン (賛否) で報道されているのかをまとめたデータベースである。

本論文では、NSContrast と GDELT についての簡単な紹介を行ったあと、NSContrast により GDELT データを分析した事例について紹介する。

*連絡先： 北海道大学大学院情報科学研究科
〒 064-0806 札幌市北区北 14 条西 9 丁目
E-mail: yoshioka@ist.hokudai.ac.jp

[1]http://gdeltproject.org/

2 NSContrast における GDELT データの利用

2.1 NSContrast

NSContrast[1] は、日本語で記述されている各国のニューサイトとそれぞれの国の言語のニュースサイト (韓国：朝鮮日報、中国：新華社) から獲得した記事を機械翻訳したものをニュース記事データとして利用し、各国の興味の違いなどの分析を行うシステムである。

本システムでは、各記事データから、ChaSen による解析結果で得られた各記事に含まれる単語の一覧、CaboCha による固有名詞抽出の結果得られた人名・地名・組織名と Wikipedia のエントリーとの比較結果による Wikipedia 項目、意見分析システム [4] による、賛成、反対の意見情報が付加されている。

NSContrast では、この新聞記事データベースに対して、以下の機能を提供している。

1. 国別注目単語比較
   日中韓米の各国の新聞記事を対象に、バースト解析 [5] を行うことにより、各国ごとに、注目されている特徴的なキーワードと注目期間を表示する。
2. 今日の話題
   毎日の記事をクラスタリングすることにより、その日の新聞記事で多く報道されているトピックに相当する記事群を抽出する。
3. 記事検索
   単純に、キーワードを含むかどうかという検索に加え、記事に付加されている人名・地名・組織名・Wikipedia のエントリーを含むかどうかという形の問い合わせが可能。
4. 関連語の関係性分析
   与えられたキーワードを含む記事群に特徴的に存在する関連語 (共起語) とその関係を可視化する。可視化の対象としては、記事中の語句をそのまま使うだけでなく、記事に付加されている人名・地名・組織名・Wikipedia のエントリーに限定して表示させることも可能である。

   本システムにおいて、特徴的に存在する共起語とは、各国の記事において、記事群に対して相関性の高い共起語だけでなく、他の国と比べて相関性の変化が大きな語も候補とする。また、これらの共起語の内、さらに、ダイス係数によって、一定以上の共起性を持っている共起語の間にリンクを設定し、共起語グラフとして可視化する。
5. 関連語の時間遷移分析
   与えられたキーワードを含む記事群を 3 日単位で分割し、各々の記事群に対して、特徴的に存在する関連語 (共起語) を抽出し表示する。

   本機能において、特徴的に存在する共起語とは、特定の期間の記事において、記事群に対して相関性の高い共起語だけでなく、他の期間と比べて相関性の変化が大きな語も候補とする。
6. 国別比較
   入力したキーワードを含む記事数の各国ごとの遷移がグラフで表示されると共に、バースト期間を可視化する。また、対照分析の結果として、関連語の関係性分析の可視化結果のグラフを表示する。
7. 多観点分析 (マルチファセット分析)
   特定の検索語を含む記事に含まれるデータをいろいろな観点 (ファセット) から可視化する。本システムでは、以下の観点が利用可能である。

   - キーワード
   - Wikipedia のエントリー語
   - 人名
   - 組織名
   - 地名
   - 賛否

   また、各観点ごとのデータは、以下の様な形でまとめて表示する。

   - 時系列グラフ
   - 円グラフ
   - 棒グラフ
   - (重みつき) 頻度の表

   ユーザは、複数存在するグラフ表示領域毎に、表示させたい観点、表示領域に対して与える絞り込み条件や種類を設定することが可能であり、以下のような分析を支援する。

   - 複数の観点を組み合わせた検索条件の設定
     複数の観点を組み合わせた検索条件を設定すると共に、その内容を可視化する。
   - 複数の国の比較
     比較したい国に対応する形で複数のグラフを表示する。表示内容 (例えば、人名) などは、全体で固定し、各グラフに対して、国名を絞り込み条件として与えると、国ごとの違いを並べて見ることが可能となる。

2.2 Global Database of Events, Language and Tone (GDELT)

Global Database of Events, Language and Tone (GDELT) [3] は、Google News、BBC Newswire などから獲得した様々なニュースサイトの記事データから構築された非常に大規模なデータベースで、以下の 2 つのデータ形式でのデータ配布が行われている。

- GDELT Event Database
  GDELT の基本データベースで、世界中で起こった様々なイベントに関するデータベースである。イベントは、同一性の判定が行われてイベント ID が付与される。各イベントについては、イベントの起きた日時、同一のイベントを報告している記事の件数やその記事の URL、報道されている期間などの情報が記録されている。さらに、詳細な情報として、 Conflict and Mediation Event Observation (CAMEO) コード[2]に基づいたイベ

[2] http://eventdata.parusanalytics.com/data.dir/cameo.html

ントの種類 (紛争、協力…)、関連する人物・組織の情報 (国、宗教、民族、名前…) の情報や、地理情報 (地名や正規化した地名に対応する地名コード)、トーン (賛否) の数値 (正の値が賛成、負の値が反対を示し、その絶対値が、賛成や反対の度合いを示す) の平均などの情報が付与されている。2013 年 3 月以前については、1ヶ月単位や 1 年単位でのデータが作成されており、2013 年 4 月 1 日以降は、1 日単位でデータの作成が行われている。

- GDELT Global Knowledge Graph (GKG)
  GDELT Event Database におけるイベントに関するより詳細なデータベースで、主に、記事を単位として作成されており、ほぼ同じ内容を述べている記事数、記事中で述べられているイベント ID、記事の分類コード、記事中に含まれる地名、人名、組織名などのリスト、トーン (賛否) の数値 (正の値が賛成、負の値が反対を示し、その絶対値が、賛成や反対の度合いを示す)、記事の URL の情報が付与されている。このデータについては、2013 年 4 月 1 日以降は、1 日単位でデータの作成が行われている。

現在、これらのデータには、GDELT の公式サイトや Google BigQuery などを通じてアクセスすることが可能である。

2.3 GDELT データの利用

NSContrast においては、記事を単位としたデータの分析を行うため、GDELT の GDELT Global Knowledge Graph (GKG) を用いた分析を行うこととした。

多少のデータフォーマットの違いはあるものの、基本的なデータ項目の対応がとれていることから、次のような基準でデータの変換を行った。

日付 GDELT GKG が提供している日付をそのまま利用

人名、組織名、地名 GDELT GKG が提供しているデータをそのまま利用

賛否 NSContrast では、賛否を賛成、反対、中立という 3 つのレベルで分析している。これに対し GKG においてトーン (賛否) の数値は、-100(完全に反対) から 100(完全に賛成) の値をとる。よって、この値を離散化して、3 つの値に分類する必要がある。本論文では、いくつかの記事データを見ながら、実験的に、-1 より小さい場合を否定、-1〜1 を中立、1 より大きい場合を肯定と分類することとした。

サイト名 サイト名は URL の情報から抽出する。ただし、複数のサイトが同一の内容を報道していた場合には、サイトの数に対応する形で、記事のデータを複製し、データとして格納する。

サイトの所属する国 NSContrast では、国ごとの報道の違いを分析するため、各ニュースサイトを所属する国に対応づける必要がある。しかし、GDELT で扱っているサイト数の数は膨大であり、手作業で国名を割り振ることは困難であった。また、ドメイン名を利用する方法についても検討したが、多くの国で、.com や.net といったドメインが用いられており、ドメイン名にのみ依拠するのも問題であると考えられた。そのため、次のような手順で、国名の推定を行った。

1. 各国のニュースサイトの一覧を提供しているサイトである world-newspapers.com[3]を利用し、そのサイトの分類を利用して、サイト名とその所属する国のデータベースを作成した。例外として、BBC Newswire については、このリストとは別に、イギリスと判定することとした。このデータベースに存在するサイトについては、このデータを利用して所属する国名を付与する。
2. 上記のデータベースに含まれていないサイトの内、国別ドメイン (.jp,.uk など) をトップドメインに持つ場合には、そのドメイン名に対応する国名を付与する。
3. 上記の手段で判定できない場合は、所属する国名の情報を空白とする。

サイトの所属する地域 国名では、比較分析を行う単位が小さいと考えられたため、全ての国を、アメリカ合衆国、北アメリカ (アメリカ合衆国を除く)、南アメリカ、アジア、ヨーロッパ、中東、アフリカ、オセアニア、アフリカの 8 つの地域に分類し、国名が付与された場合には、対応する地域名を付与した。

URL GDELT GKG が提供している URL をそのまま利用する。ただし、同一サイトに複数の記事が存在するデータについては、そのほとんどが、CGI パラメータのみが違い、最終的に同じ内容の記事が表示される URL であったため、同一サイトの記事については、URL 毎に記事を分割するのではなく、最も短い URL をその記事を代表する URL として利用する。

[3] http://www.world-newspapers.com/

これらの変換を行うことにより、新聞記事の本文やタイトルに関する情報以外については、NSContrast とほぼ同等のデータを作成することが可能となった。

3 分析事例

3.1 記事データベースの構築

本論文では、GDELT GKG データの 2013 年 4 月 1 日〜2013 年 12 月 25 日のデータを用いて、分析実験を行った。2.3 節で述べたデータ変換を行うことにより、11,177,775 件の記事を得た。表に各地域ごとの記事数を示す。今回、提案している国名の付与のアルゴリズムは、まだ十分でないため、約 38%(4,280,168) の記事について、国名の推定を行うことができなかった。

表 1: 各地域ごとの記事数

アメリカ合衆国	2,933,282	アジア	1,295,274
ヨーロッパ	1,258,470	中東	343,595
アフリカ	392,768	オセアニア	384,462
北アメリカ	254,204	南アメリカ	35,552
分類不能	4,280,168		

3.2 分析事例

まず、NSContrast の国別注目単語比較の機能を用いることにより、それぞれの地域で注目されている単語や固有名詞間の比較を行うことが可能となる。図 1 は、2013 年 9 月 29 日における人名に対する比較で、バースト度の高い順に、名前が表示される。字の大きさは、各地域での相対的な記事数 (全体の記事のうち、固有名詞を含む記事のる割合) を表している。また、背景の赤色は、その人名が特定の地域のみで注目が高いことを示している。

赤色は特定の地域のみで注目されている人名で、最も色の濃いところが、1 つの地域のみで注目されていることを示し、色が薄くなるほど、注目されている地域が増えることを示している。

また、この注目度については、1 列目は、話題の継続期間を考慮したもので、2 列目は、話題の継続期間を考慮しないものである。継続期間を考慮しない場合には、対象とする日付の各単語のバースト度によりランキングを行い、継続期間を考慮する場合には、単語が継続してバーストしている場合に、バースト開始時からの各々のバースト度を加算したものを用いてランキングを行う。

この表から、「ronald bechtold (Mr. Ronald Bechtold:アメリカ国防総省の主席情報官)」が、世界中で注目されていることが分かる。また、同日、「nawaz sharif(Mr. Nawaz Sharif:パキスタンの首相)」がアメリカで 10 位以内の注目されたことが分かる。

「nawaz sharif」の名前をクリックすると、「nawaz sharif」に関する注目度の情報ウィンドウが表示される (図 2)。このウィンドウには、各国での注目度の順位、対応する記事数、注目されている期間などが表示される。赤い文字で示された期間は、その期間において、継続期間を考慮した注目度が 10 位以内に入ったことがある期間を示している。

この表から、アジアでは、6 月頃に、10 位以内に入る程度に注目されていたことが分かる。

また、関連語の関係性分析を行うことにより、「nawaz sharif」を含む記事に特徴的に現れる他の人物を表示することができる (図 3)。赤いノードは、他の地域では、注目されていないが、特定の地域では、比較的注目されている人物を示している。

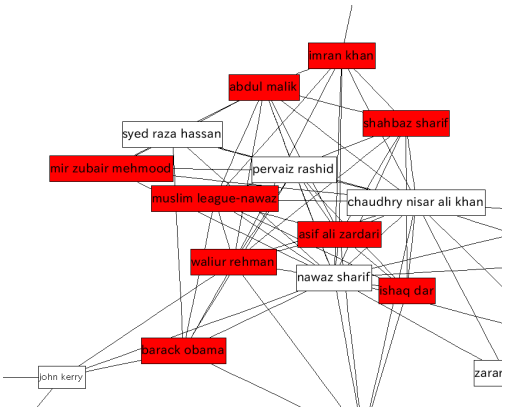


図 3: Results of the term collocation analysis

「imran khan(Mr. Imran Khan: パキスタンの政治家で、元クリケットの選手)」は、他の地域に比べて、アジアのニュース記事において、良く、「nawaz sharif」と同じ記事に登場することが分かる。この様に、特定の地域に限定した分析などを行うことで、世界的には知られていないが、特定の地域では、著名な人物などを見つけ出すことが可能となる。

最後に、この「nawaz sharif」に対して、多観点分析を行う。図 4 は、多観点分析のインターフェースで、左半分のウィンドーは、対象とする記事を観点ごとに絞り込むための検索式を作成するための領域である。この図では、対象となる人物が、「nawaz sharif」で、分析対象とする期間を「2013 年 4 月 1 日から」に設定している。

右側のグラフは、時系列グラフであり、Graph1(左上) が地域、Graph2(右上) が国名、Graph3(左下) が人名、Graph4(右下) が賛否を表している。このグラフから、「nawaz sharif」は、主に、アジア、特に、パキスタンとインドの記事に多く現れていることが確認できる。

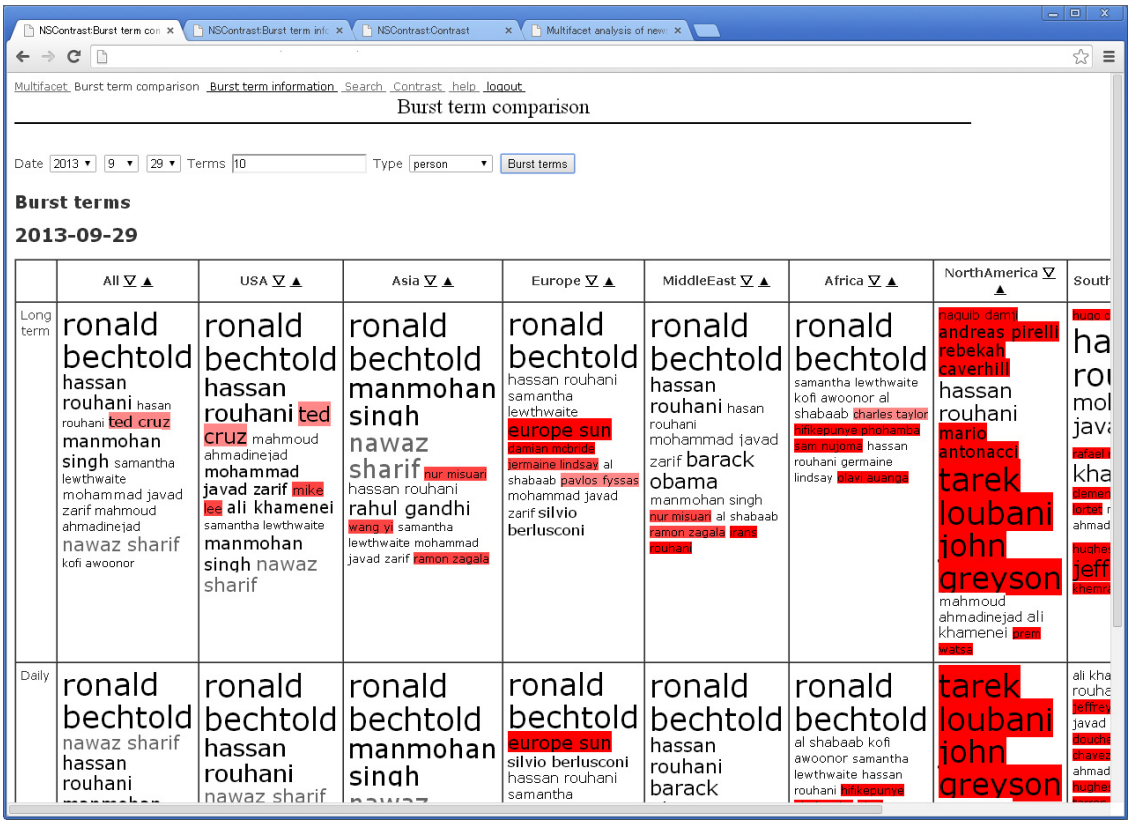
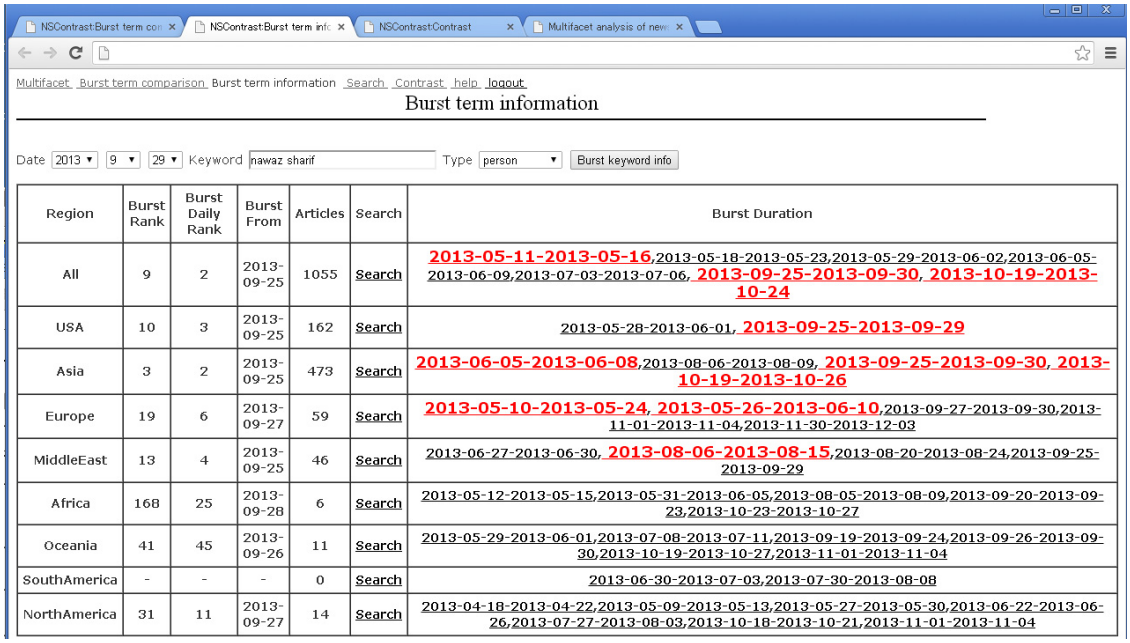


図 1: 国別注目単語比較



Region	Burst Rank	Burst Daily Rank	Burst From	Articles	Search	Burst Duration
All	9	2	2013-09-25	1055	Search	2013-05-11-2013-05-16,2013-05-18-2013-05-23,2013-05-29-2013-06-02,2013-06-05-2013-06-09,2013-07-03-2013-07-06, 2013-09-25-2013-09-30, 2013-10-19-2013-10-24
USA	10	3	2013-09-25	162	Search	2013-05-28-2013-06-01, 2013-09-25-2013-09-29
Asia	3	2	2013-09-25	473	Search	2013-06-05-2013-06-08,2013-08-06-2013-08-09, 2013-09-25-2013-09-30, 2013-10-19-2013-10-26
Europe	19	6	2013-09-27	59	Search	2013-05-10-2013-05-24, 2013-05-26-2013-06-10,2013-09-27-2013-09-30,2013-11-01-2013-11-04,2013-11-30-2013-12-03
MiddleEast	13	4	2013-09-25	46	Search	2013-06-27-2013-06-30, 2013-08-06-2013-08-15,2013-08-20-2013-08-24,2013-09-25-2013-09-29
Africa	168	25	2013-09-28	6	Search	2013-05-12-2013-05-15,2013-05-31-2013-06-05,2013-08-05-2013-08-09,2013-09-20-2013-09-23,2013-10-23-2013-10-27
Oceania	41	45	2013-09-26	11	Search	2013-05-29-2013-06-01,2013-07-08-2013-07-11,2013-09-19-2013-09-24,2013-09-26-2013-09-30,2013-10-19-2013-10-27,2013-11-01-2013-11-04
SouthAmerica	-	-	-	0	Search	2013-06-30-2013-07-03,2013-07-30-2013-08-08
NorthAmerica	31	11	2013-09-27	14	Search	2013-04-18-2013-04-22,2013-05-09-2013-05-13,2013-05-27-2013-05-30,2013-06-22-2013-06-26,2013-07-27-2013-08-03,2013-10-18-2013-10-21,2013-11-01-2013-11-04

図 2: Nawaz Sharif 首相に対する注目度の情報

図 4: Nawaz Sharif を対象とした多観点分析

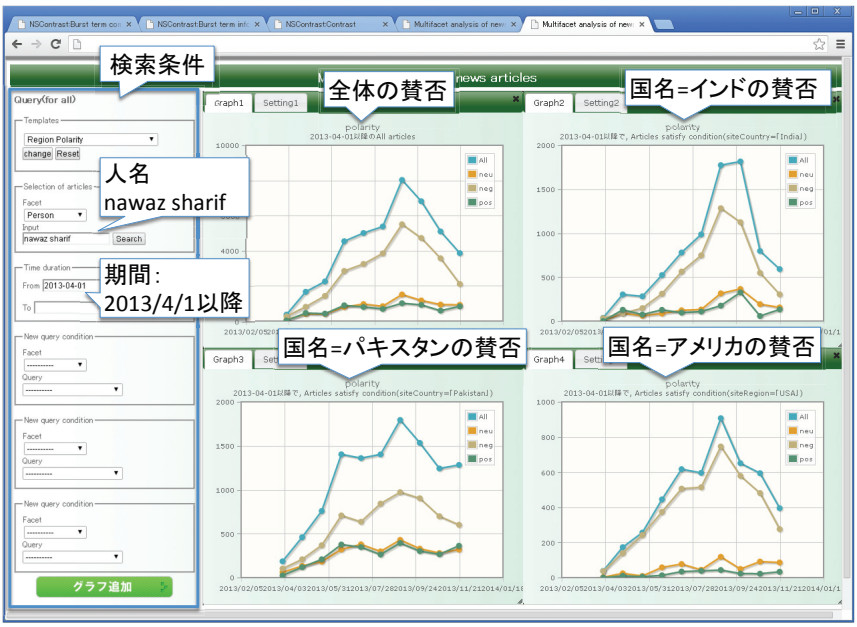


図 5: Nawaz Sharif と Imran Khan を対象とした多観点分析

図 6: Nawaz Sharif の賛否に関する比較

この分析ウィンドーにおいて、検索式を修正すると、条件を満たした対象記事に対して、グラフが更新される。図 5 は、検索式に、先ほどの分析で得た人物「imran khan」を追加した場合を示す。このグラフから、6 月の段階では、「imran khan」は、アジアとヨーロッパでは、それなりに注目されていたが、アメリカでは、あまり注目されていないことが分かる。しかし、11 月になると、この「imran khan」は、「アメリカの無人攻撃機を利用したタリバンへの攻撃」に反対することで、アメリカでも注目されることになる。この様に、アメリカのニュースサイトを見ていただけでは、突然、注目された人物という事になるが、NSContrast のように、世界のニュースサイトを比較分析していくシステムを用いることにより、この様な自体に対しても、より多角的な分析ができるようになると考えている。

また、この多観点分析システムは、様々な観点に関する情報を並べて提示するだけでなく、異なる条件の検索結果の比較を行うことも可能である。図 6 は、「nawaz sharif」を含む記事における賛否に対する様々な国の比較を行った結果である。Graph1(左上) が全てのニュースサイト (追加検索条件なし)、Graph2(右上) がインド (サイト国＝インド)、Graph3(左下) がパキスタン (サイト国＝パキスタン)、Graph4(右下) がアメリカ (サイト国＝アメリカ) の賛否の時間遷移を表示したグラフとなる。この比較から、アメリカとパキスタンにおける賛否の違いなどを読み取ることが可能となる。

3.3 考察

この分析の結果、NSContrast が提供する比較分析機能は、世界中では注目されてるとはいえないが、特定の地域では注目されているような、ローカルな情報を見つけ出すのに有用であると考えている。

しかし、より精緻な分析を行うためには、以下の点において、データを充実させる必要があると考えている。

- サイトの所属する国の判定
  現時点では、38%の記事について、サイトの所属する国が正しく判定されていない。これについては、より適切な情報となるように、更新することが望ましい。

- ニュース記事のタイトル
  GDELT の GKG データは、記事を単位として作られているが、記事から抽出された情報をリストとしてみるだけでは、閲覧性が高くない。少なくとも、記事のタイトルを、別のデータとして作成した上で、表示させることが出来ると、閲覧性の向上につながると考えている。

4 まとめ

本論文では、ニュース記事の多観点分析を支援する NSContrast と、ニュース記事を基盤とした大規模なイ

ベント情報である GDELT について紹介を行い、このデータを NSContrast で用いる方法について述べた。これまでの自前で作るデータベースに比較して、非常に大規模なデータが使えるという利点はあるが、こちらの目的に合わせてデータを変換する必要があったり、不足しているデータがあると言った問題点があった。

しかし、大規模なデータを使うことで、より、実際的な分析が可能になると考えられるので、このデータの利用方法について、さらなる検討を行っていきたいと考えている。

謝辞

本研究の一部は、科研費基盤研究 (B) 25280035 により行われた。ここに記して、謝意をあらわす。

参考文献

[1] 吉岡真治, 神門典子, 関洋平. 複数国の新聞サイトを比較分析する nscontrast の実験的分析. 情報処理学会デジタルドキュメント研究会, 2011-IFAT-103, 2011. IFAT-103-2.

[2] Masaharu Yoshioka and Noriko Kando. Multi-faceted analysis of news articles by using semantic annotated information. In *Proceedings of the fifth workshop on Exploiting semantic annotations in information retrieval*, ESAIR '12, pp. 19–20, New York, NY, USA, 2012. ACM.

[3] Kalev Leetaru and Philip A. Schrodt. Gdelt:global data on events, location, and tone, 1979-2012. In *ISA Annual Convention 2013*, Vol. 2, p. 4, 2013.

[4] Yohei Seki, Noriko Kando, and Masaki Aono. Multilingual opinion holder identification using author and authority viewpoints. *Information Processing & Management*, Vol. 45, No. 2, pp. 189–199, 2009.

[5] Jon Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 91–101, New York, NY, USA, 2002. ACM Press.