語の共起のバースト検出に基づく研究トレンドの可視化

桂井麻里衣† 小野 峻佑††

† 同志社大学理工学部情報システムデザイン学科 〒610–0394 京都府京田辺市多々羅都谷 1-3
†† 東京工業大学科学技術創成研究院未来産業技術研究所 〒226–8503 神奈川県横浜市緑区長津田町 4259
E-mail: †katsurai@mm.doshisha.ac.jp, ††ono@isl.titech.ac.jp

あらまし 科学技術の動向把握や異分野共同研究などの企画戦略の支援には，各研究分野における萌芽的技術の自動検出やトピック変遷の可視化が重要とされている．本論文では，これら二つを同時実現する新たなマッピング手法として，語の共起のバースト検出に基づく研究トレンドの可視化を提案する．はじめに，対象とする分野の学術論文を一定の時間幅で分割し，各時間区間において語の共起頻度を算出する．次に，共起頻度の時系列からなる行列を，バーストに強く関連する急激な変化を要素に含むスパース行列と，残りの定常的な成分で構成されるスムース行列に分解する．最後に，得られた行列の非ゼロかつ正の要素をエッジ重みとするネットワークを構築することで，対応する時間区間の研究トレンドを可視化する．本文の最後には，人工データを用いた実験と実データを用いた実験を行い，提案手法の有効性を示す．

キーワード 研究トレンド，共起語，バースト検出，サイエンスマッピング

1. はじめに

国際会議・論文誌の継続的増加やプレプリントサーバの普及，研究成果のオープン化に伴い，膨大な量の学術情報が日々蓄積されている．研究企画戦略に携わる者はタイムリーに研究動向を把握する必要があるが，全ての論文に目を通すことは難しい．また昨今では異分野融合による研究開発が推進されており，各々の専門分野のパラダイムや歴史の共有が求められる．このような研究分野に関する理解の促進に向け，注目技術の自動検出や研究トピック変遷の可視化が活発に研究されている [1–12].

従来研究では，分野の知識構造を概念空間にマッピングするために，学術論文集合における語の共起関係 [1] を用いる．まず論文集合を一定時間幅で分割し，時間区間ごとに単語ペアの共起頻度を算出する．次に，単語をノード，語の共起関係をエッジ，共起頻度をエッジ重みにもつ共起語ネットワークを構築する．一般に，ネットワークのエッジ数は重みの閾値処理により決定される．構築した共起語ネットワークを時系列に並べることで，研究分野の発展を分析する [2–8]. しかしこのアプローチでは，定常的な語の共起と，急激に増加した語の共起を識別することができない．前者の存在は，各時間区間を特徴付ける研究トピックの理解や萌芽的技術の発見を妨げる要因となる．そこで我々は，直前の時間区間から急激に重みが増加したエッジの検出に着目し，概念空間において萌芽的技術の変遷マッピングを実現する．

本論文では，語の共起の急激な増加（バースト）の検出に基づく研究トレンドの可視化を提案する．提案手法は，まず各時間区間で全ての単語ペアの共起頻度を算出し，それらを時系列に並べた行列を構築する．次に，行列を急激な時間変化を要素として含むスパース行列と，定常的な要素からなるスムース行列の二つに分解する．我々はこの最適化問題を高速スパース－スムース行列分解と呼ぶ．得られたスパース行列の非ゼロかつ正の要素から共起語ネットワークを再構築することで，各時間区間のバーストに相当する研究トピックが可視化される．提案手法の有効性を評価するために，まず人工データを用いてベースライン手法とバースト検出性能を比較する．次に，人工知能，コンピュータビジョン，情報通信をスコープとする国際会議を対象とし，過去 20 年分の論文を用いてそれぞれの研究トレンドを可視化する．さらに，提案手法の応用として，各国際会議の発展スピードを表す指標を導出する．

本論文の主な貢献は下記の通りである．

- 研究分野における萌芽的技術の発見とその変遷の可視化を目的とし，語の共起のバースト検出手法を新たに提案する．提案手法は時系列共起語ネットワークと単一パラメータのみを入力とするよう設計してあり，様々な分野のサイエンスマッピングに新たな知見を提供できる．
- 人工データを用いた定量評価と実データを用いた定性評価により提案手法の有効性を示す．特に実データを用いた実験では，提案手法から研究分野の発展スピードを表す指標を導出し，応用可能性を検討する．

本文の構成は以下の通りである．まず，2 章で本研究の関連研究を説明する．3 章では時系列共起語グラフの高速スパース－スムース行列分解に基づくバースト検出手法を提案する．提案手法の有効性を評価するために，4 章では人工データにおけるバースト検出性能を報告し，5 章では三つの国際会議をケーススタディとしたトレンド検出結果を提示する．最後に，6 章において本文をまとめ，今後の研究課題について述べる．

2. 関連研究

学術情報のビッグデータ化が進む中，研究者やリサーチ・アドミニストレータ，政策立案者らの活動を支援するには，各分野の知識構造の抽出・可視化が必要不可欠である．これまで，学

術情報の構成要素である論文，研究者，研究用語の関係性を表す複雑ネットワークの分析方法が種々提案されてきた．例えば研究者の共著関係ネットワークは，影響力の強い研究者や研究グループの発見に有用である [13]．また，論文の引用関係ネットワークは，分野の類似性の発見のみならず，研究の潮流の俯瞰につながるといわれている [14]．しかし，論文は引用されるまでに時間を要するため，最新の動向把握は難しい．

一方，論文テキストを情報源とする共起語ネットワークは，学術論文がまとめて公開された時点で構築可能であるため，萌芽的技術の早期発見に適している．加えて，技術名や技術間の関連性を概念空間で直接表現できるという利点がある．語と語の間の意味的なつながりの図示は，知識構造の直感的な理解を促進するといわれている [15–17]．このような利便性から，近年も科学教育 [2] や経営戦略論 [4]，看護研究 [5]，情報検索 [6] など様々な分野の発展分析に用いられている．分析対象の粒度を細かくし，特定の学術論文誌や国際会議で出版された論文集合に適用した例も存在する [9,10]．一般に，ネットワークのエッジ数は共起頻度の閾値処理で決定されるが，多くの分析事例において定常的な語の共起が大部分を占める傾向にある．

近年，時系列共起語ネットワークを用いた研究トピック変遷の可視化手法 [11,12] が提案された．これらの手法は，各時間区間のネットワークにクラスタリングを適用し，語のクラスタを時間追跡することでトピックの変遷を可視化する．可視化にはクラスタ内の最頻出語が用いられるため，出力されるマップは本研究で目的とする萌芽的技術の発見には適さない．対象を学術論文に限らない場合，テキスト集合におけるトピック遷移を可視化する手法として，確率的トピックモデルの時間拡張（例：Dynamic Topic Model [18]）も検討されている．これらの研究においても，トピック生成の過程で頻出語が重視される傾向にある．我々が提案する手法は単語対ごとに共起頻度の定常性を考慮するため，語のクラスタリングやトピックモデルに比べ，特徴的なテクニカルタームの抽出に適している．

語のバースト検出は，テキストストリームからのトピックマイニングやモニタリングなどに応用されている．代表的なアプローチとして，事象の発生頻度の閾値処理（例：移動平均法の利用 [19]）や，状態遷移のモデル化（例：Kleinberg の有限オートマトンに基づく手法 [20]）が挙げられる．現在公開されている学術情報分析ツール [21] では，Kleinberg の手法に基づく語のバースト検出機能が実装されている．本研究では，単一の語ではなく語の共起に着目することで萌芽的技術を説明するためのネットワークを構築する．また，利便性の高いマッピング技術とするため，時系列共起語ネットワークと単一パラメータのみを入力とするよう定式化した．

3. 提 案 手 法

本章では，語の共起のバースト検出に基づく研究トレンドの可視化手法を提案する．提案手法の概要を図 1 に示す．まず，各時間区間で共起語ネットワークを構築し，エッジ重みの時系列から行列を算出する（3.1 節）．次に，行列を滑らかな時間

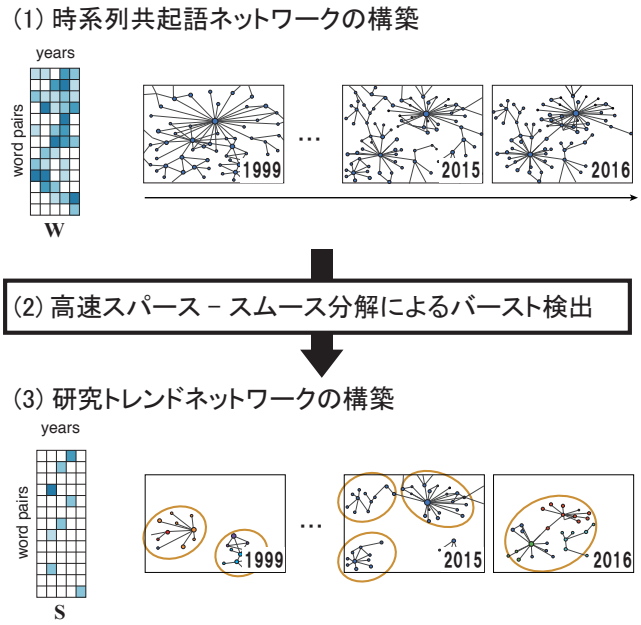


図 1　提案手法の概要．入力された時系列共起語ネットワークに対し，エッジ重みのバーストを検出し，研究トレンドを表すネットワークを出力する．

変化と急激な時間変化からなる二つの行列に分解する最適化問題を解く（3.2 節）．最後に，得られたスパース行列から時系列共起語ネットワークを再構築する（3.3 節）．

3.1 語の共起確率の算出

対象分野の論文集合を Ω で表し，Ω 中のユニークな単語の数を M とおく．各論文の出版日時に基づき，論文集合 Ω を一定時間幅で分割し，連続した部分集合 $\Omega(1), \Omega(2), \cdots, \Omega(T)$ ($\Omega = \cup_{t=1}^{T} \Omega(t)$) を得る．次に，$t$ ($t \in \{1, 2, \cdots, T\}$) 番目の時間区間において共起語ネットワークを構築する．共起頻度に基づくエッジ重みの算出には様々な方法があるが，本文では文献 [10] と同様，二つの単語の共起回数を論文数で正規化した値をエッジ重みに用いる．

$$\tilde{W}_t(i,j) = \frac{n_t(i,j)}{|\Omega(t)|}, i \neq j, i, j \in \{1, 2, \cdots, M\}, \qquad (1)$$

ここで，$n_t(i,j)$ は i 番目と j 番目の単語が t 番目の時間区間で共起した回数を表し，$|\Omega(t)|$ は集合 $\Omega(t)$ の論文数を表す．式 (1) を全ての単語ペアについて算出すると，対称行列 $\tilde{\mathbf{W}}_t \in \mathbb{R}^{M \times M}$ を得る．これが従来の共起語ネットワークの隣接行列に相当する．

提案手法では，全ての単語ペアのエッジ重み（つまり行列 $\tilde{\mathbf{W}}_t$ の上三角成分）を要素にもつ列ベクトル $\mathbf{w}_t \in \mathbb{R}^N$ ($N = M(M-1)/2$) を算出する．この列ベクトルを時系列に並べることで行列 $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_T] \in \mathbb{R}^{N \times T}$ を構築する．

3.2 高速スパースースムース行列分解

本節では，時系列のエッジ重みからバーストを検出するための行列分解手法を提案する．具体的には，次のような最適化問題を解くことで，前節で算出した行列 $\mathbf{W}$ を，バーストに強く関連する急激な変化を要素として含むスパース行列と，残りの

定常的な成分で構成されるスムース行列に分解する．

$$\min_{\mathbf{S}} \frac{1}{2}\|D(\mathbf{W}-\mathbf{S})\|_F^2 + \lambda\|\mathbf{S}\|_1, \tag{2}$$

ここで，$\mathbf{S}$ はスパースな行列，D は隣り合う列同士の差分を計算する線形作用素，$\|\cdot\|_1$ と $\|\cdot\|_F$ はそれぞれ ℓ_1 ノルムとフロベニウスノルムを表す．問題 (2) の第二項は $\mathbf{S}$ のスパース性を促進する項である．(注1) 第一項はバースト成分を除いた行列 $\mathbf{W}-\mathbf{S}$ が列方向（時間方向）に滑らかになるように働く．

問題 (2) は微分不可能な項を含む凸最適化問題であるが，fast iterative shrinkage-thresholding algorithm (FISTA) [22] と呼ばれるアルゴリズムを応用することで高速に解を求めることができる．提案手法の詳細に立ち入る前に，FISTA の一般型について説明する．微分可能な凸関数 f（勾配 ∇f がリプシッツ連続であると仮定）と近接写像 [23](注2) が効率的に計算可能な凸関数 g に対して，次のような最適化問題を考える．

$$\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}), \tag{3}$$

この問題の解を，FISTA は次のようなアルゴリズムで求める：任意の初期ベクトル $\mathbf{y}_0$ と $z_0 = 1$ に対し，以下を反復する．

$$\begin{cases} \mathbf{x}_{k+1} := \mathrm{prox}_{\frac{1}{L}g}(\mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k)), \\ z_{k+1} := \frac{1+\sqrt{1+4z_k^2}}{2}, \\ \mathbf{y}_{k+1} := \mathbf{x}_k + \frac{z_k-1}{z_{k+1}}(\mathbf{x}_k - \mathbf{x}_{k-1}), \end{cases} \tag{4}$$

ここで k は反復数，L は ∇f のリプシッツ定数である．

FISTA を問題 (2) に適用していこう．問題 (3) において，$f(\mathbf{S}) := \frac{1}{2}\|D(\mathbf{W}-\mathbf{S})\|_F^2$，$g(\mathbf{S}) := \lambda\|\mathbf{S}\|_1$ と定義すると，明らかに f は微分可能であり，その勾配は

$$\nabla f(\mathbf{S}) = -\mathbf{D}^\top\mathbf{D}(\mathbf{W}-\mathbf{S})$$

で与えられ，リプシッツ定数 L は $\|D\|_{op}^2$ となる（$\|\cdot\|_{op}$ は作用素ノルム）．一方，g は ℓ_1 ノルムであるため，その近接写像はよく知られたソフト閾値値処理に帰着される．

$$[\mathrm{prox}_{\frac{1}{L}g}(\mathbf{S})]_{i,j} = \mathrm{sgn}(S_{i,j})\max\{S_{i,j} - \tfrac{\lambda}{L}, 0\}. \tag{5}$$

結果として問題 (2) は FISTA に基づく Algorithm 1 で解くことができる．

最後に，FISTA が問題 (2) を解くためのアルゴリズムとして妥当である理由を説明する．第一に，FISTA が非常に高速なアルゴリズムであることが挙げられる．具体的には，その収束レートが $\mathcal{O}(1/k^2)$ [22] であり，これは勾配に基づくあらゆるアルゴリズムの中で（一種の）最適なレートであることが知られている．第二に，計算量が小さいことが挙げられる．既存の行列分解手法（例えばロバスト主成分分析 [24]）では，各反復で特異値分解が必要となるため，行列サイズが大きい場合に計算が困難になってしまう．一方，FISTA に基づく提案アルゴリズムの各ステップの評価（勾配計算とソフト閾値処理）は行列の要素数に対して線形なオーダーで済むためスケーラブルである．

(注1)：ℓ_1 ノルムは非ゼロ要素数をカウントする非凸関数である ℓ_0 擬ノルムの凸緩和になっており，スパース性を評価する指標として広く用いられている．

(注2)：任意の $\gamma > 0$ に対し，下半連続な真凸関数 h の近接写像は $\mathrm{prox}_{\gamma h}(\mathbf{x}) := \arg\min_{\mathbf{y}} h(\mathbf{y}) + \frac{1}{2\gamma}\|\mathbf{y}-\mathbf{x}\|^2$ で与えられる．

Algorithm 1: FISTA による行列の高速スパース/スムース分解

input : $\mathbf{Y}_0 = \mathbf{W}$, $z_0 = 1$, $L = \|D\|_{op}^2$, λ, $k = 0$

while *A stopping criterion is not satisfied* **do**

  $\mathbf{S}_{k+1} = \mathbf{Y}_k + \frac{1}{L}D^*D(\mathbf{W} - \mathbf{Y}_k)$;

  $\mathbf{S}_{k+1} \leftarrow$ soft-thresholding$(\mathbf{S}_k, \frac{\lambda}{L})$ by (5);

  $z_{k+1} := \frac{1+\sqrt{1+4z_k^2}}{2}$;

  $\mathbf{Y}_{k+1} := \mathbf{S}_k + \frac{z_k-1}{z_{k+1}}(\mathbf{S}_k - \mathbf{S}_{k-1})$;

  $k \leftarrow k+1$;

output: $\mathbf{S}_k$

3.3 研究トレンドネットワークの構築

最後に，問題 (2) を解いて得られたスパース行列 $\mathbf{S} \in \mathbb{R}^{N\times T}$ から，研究トレンドを可視化する．まず，$\mathbf{S}$ の t 列目のベクトルを対称行列 $\tilde{\mathbf{S}}_t \in \mathbb{R}^{M\times M}$ に変換する．このとき，$\tilde{\mathbf{S}}_t$ の非ゼロ要素が正のとき，対応する単語ペアの発生頻度は t 番目の時間区間でバーストに相当する．よって，負の非ゼロ要素を 0 に置き換えた $\tilde{\mathbf{S}}_t$ を隣接行列としてネットワークを再構築すればよい．

構築した研究トレンドネットワークに対し，Louvain 法 [25] を適用し，語のコミュニティを発見する．同一コミュニティに属するノードには同じ色を割り当てることで，研究トピックを構成する単語の視認性を高める．

4. 人工データを用いた定量評価

研究トレンド抽出の正解データは公開されておらず，手法の性能評価が困難である．そこで本章では，人工データを用いた実験を行い，高速スムース－スパース行列分解によるバースト検出の性能を評価する．以降，4.1 節で人工的なデータセット構築手順について述べ，4.2 節でベースライン手法を説明する．最後に 4.3 節で各手法の性能を評価する．

4.1 データセット

正解データを用意するため，時系列共起語ネットワークの一部のエッジ重みについて，バーストを人為的に発生させた．以下，データセット構築手順を説明する．まず，Rouiter-21578 内のカテゴリ “trade” に属する文書を用いて 5-gram 言語モデルを構築し，10000 枚の擬似文書集合を生成する試行を $T = 50$ 回繰り返した．このとき，各文書の単語数は最大 12 に設定した．各試行 t $(t = 1, 2, \cdots, T)$ で生成された擬似文書集合を Ω_t で表す．全集合 $\{\Omega_t\}_{t=1}^{50}$ において出現回数が 30 未満となった単語を除外した結果，ユニークな語の数は $M = 6,696$ となった．

ボキャブラリ中の i, j $(i, j \in \{1, \cdots, M\}, i < j)$ 番目の単語ペアについて，集合 Ω_t における共起頻度を $G_t(i,j)$ とおく．また，時系列 $\{G_t(i,j)\}_{t=1}^T$ における共起頻度の平均を $\mu(i,j)$，標準偏差を $\sigma(i,j)$ で表す．本実験では，時間区間 t において，バースト発生対象とする単語ペア (i', j') を確率 $p = 0.01$ で選択し (注3)，元のエッジ重みを次式によりバースト値に置き換

(注3)：共起頻度の非常に少ない単語ペアについては式 (6) を用いたバースト発

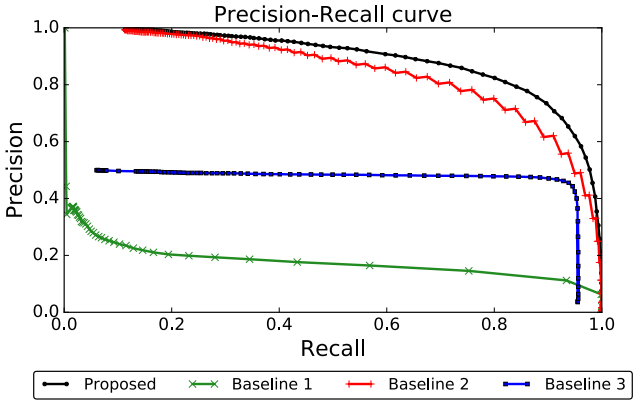


図 2　提案手法およびベースライン手法による PR 曲線.

表 1　PR 曲線の AUC.

提案手法	ベースライン 1	ベースライン 2	ベースライン 3
0.890	0.178	0.844	0.465

えた.

$$G_t(i',j') \leftarrow \mu(i',j') + u_t(i',j')\sigma(i',j'), \qquad (6)$$

ここで，$u_t(i',j')$ は一様分布 $[3,6]$ からランダムに発生させた値を表す．以上の方法で，各時間区間で一部のエッジ重みがバーストとなる時系列共起語ネットワークを得た．なお，単語ペア (i,j) に対し連続した時間区間で式 (6) が適用された場合，最初の時間区間のみをバーストとしてラベル付けした．最終的に，$t = 1, 2, \cdots, T$ でバースト重みを加えたエッジの総数は 15,592 となった.

4.2　ベースライン手法

本実験では，以下の三つのベースライン手法と提案手法の性能を比較する.

ベースライン 1（行列要素の閾値処理）：時刻 t における i, j 番目の単語について，エッジ重み $G_t(i,j)$ が閾値 τ_1 以上の場合にバーストとみなす．共起語ネットワークのエッジ数決定によく用いられるアプローチである.

ベースライン 2（時間微分の閾値処理）：連続した時間区間におけるエッジ重みの時間微分 $G_t(i,j) - G_{t-1}(i,j)$ が閾値 τ_2 以上の場合にバーストとみなす．提案手法に最も近い手法である.

ベースライン 3（Kleinberg の手法 [20]）：各単語ペアのエッジ重みの時系列を，通常状態とバースト状態の二つの状態からなるオートマトンでモデル化する．与えられた時系列データに対し，Viterbi アルゴリズムを用いて最適な状態シーケンスを算出したあと，各時間区間におけるバーストの強さを定量化する [26]．モデルパラメータは s, γ の二種類となる（詳細は文献 [20] を参照されたい）．本実験では前者を $s = 2.0$ と固定し，γ の値を変更して性能を調査する.

生が難しいため，$\mu(i,j) < 10$ となる単語ペアを予め選択肢から除外した.

表 2　5 章の実験で用いるデータセットの詳細.

会議名	AAAI	CVPR	INFOCOM
期間	1993-2016	1994-2015	1997-2016
	(2001,2003 以外)	(2002 以外)	
論文数	6,454	6,863	5,807
語彙数	7,480	6,024	6,529

4.3　実験結果

提案手法およびベースライン手法を実験データセットに適用し，次式の Precision および Recall を算出した.

$$Precision = \frac{\text{正しくバーストとして検出できたエッジ数}}{\text{手法によりバーストとして検出したエッジ数}},$$

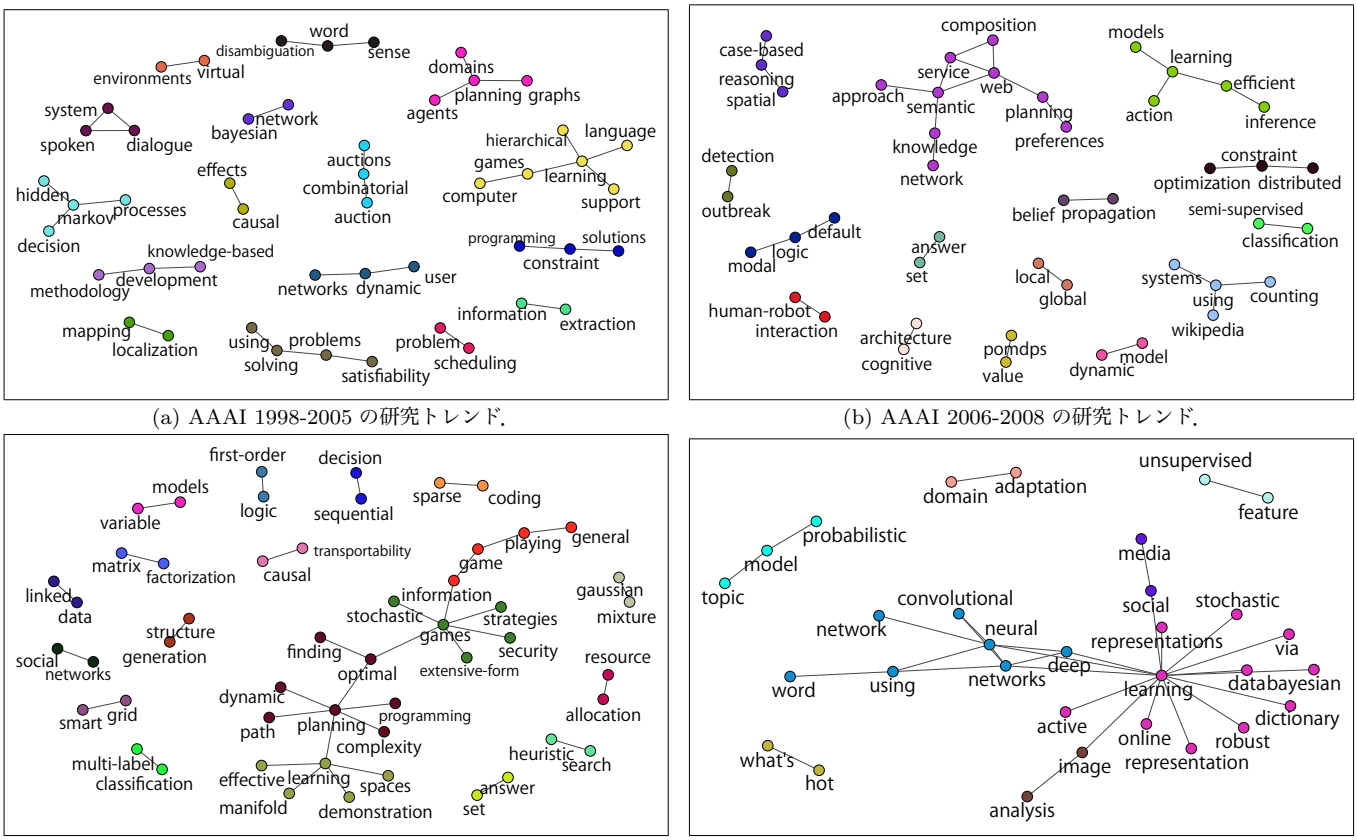
$$Recall = \frac{\text{正しくバーストとして検出できたエッジ数}}{\text{バースト重みを加えたエッジ数}}.$$

各手法による Precision-Recall (PR) 曲線を図 2，各 PR 曲線の面積（Area Under Curve; AUC) を表 1 に示す．図表より，提案手法が最も精度良くバーストを検出できたことがわかる．特にベースライン 1 は Precision が非常に低い．したがって，従来のサイエンスマッピングで用いられる単純な閾値処理では，各時間区間を特徴付けるネットワークの抽出は明らかに困難といえる．ベースライン 3 の Kleinberg のバースト検出手法は二つのパラメータを必要とするが，効率的にパラメータを自動設定する手法は提案されていない．また結果が示すとおり，本実験設定では高精度な検出が困難であった．提案手法は単一パラメータ λ のみでネットワークのスパース性を変更することが可能であり，精度と利便性ともにベースライン 3 より優れている．ベースライン 2 は時間微分を算出するため，急激な重み変化の検出を目的とする提案手法に最も近い手法である．しかし，$\mu(i,j)$ の値が大きい単語ペアについては，他の語に比べて相対的に時間微分が大きくなりバースト誤検出が増える．これに対し提案手法は単語ペアごとに時間方向の滑らかさを考慮するため，単純な時間微分の閾値処理に比べてバーストを高精度に捉えられる．以上の結果から，提案手法の有効性が確認できた.

5.　国際会議論文を用いた定性評価

本章では，提案手法を実際の論文集合に適用し，研究トレンド可視化結果を定性的に評価する．実験対象として AAAI, CVPR，INFOCOM という三つの国際会議を選択した．それぞれ人工知能，コンピュータビジョン，情報通信をスコープとするトップレベルの会議であり，各分野の最先端技術を十分カバーしていると考えられる．各会議に対し，Digital Bibliography & Library Project (DBLP)(注4) から過去 20 年分の論文タイトルを収集した．各論文タイトルの単語集合から，Natural Language Toolkit (NLTK)(注5) に収録されている英語ストップワード（例：“a”, “the”, “from”）を除去し，“:”, “?”, “!” などの不要な記号を全て削除した．データセットの期間，論文数，語彙数を表 2 に示す.

(注4)：http://dblp.uni-trier.de/
(注5)：http://www.nltk.org/



(a) AAAI 1998-2005 の研究トレンド.

(b) AAAI 2006-2008 の研究トレンド.

(c) AAAI 2010-2013 の研究トレンド.

(d) AAAI 2014-2016 の研究トレンド.

図 3　国際会議 AAAI の発展を示す研究トレンドネットワークの可視化： (a) 1998-2005, (b) 2006-2008, (c) 2010-2013, (d) 2014-2016. パラメータ λ は各ネットワークのエッジ数が 30 前後になるよう調節した.

本実験では，式 (4) のリプシッツ定数 L を全て 4.0 に設定した (注6). 各会議を $c \in \{\mathrm{AAAI}, \mathrm{CVPR}, \mathrm{INFOCOM}\}$ で表し，c のデータセットを T_c 個の部分集合へ分割した. c と時間区間の番号 $t \in \{1, 2, \cdots, T\}$ の組み合わせについて，元の共起語ネットワークのエッジ集合を $E(c,t)$，提案手法をパラメータ λ で適用して得られたエッジ集合を $\mathcal{E}(c,t,\lambda)$ で表す. 実験には Intel Xeon プロセッサ E5-1680V3 (3.1GHz) 搭載ワークステーションを用いた. $\lambda = 10^{-2}$，$c = \mathrm{AAAI}$，$T_c = 20$ としたとき，スパース－スムース行列分解の実行時間は約 0.93 秒であり，高速なバースト検出が可能であった.

以降，まず 5.1 節で AAAI のトピック発展の様子を分析し，5.2 節で CVPR と INFOCOM を例として提案手法およびベースライン 1 の結果を定性的に比較する. 最後に，提案手法の応用として，5.3 節で各国際会議の発展スピードを表す指標を導出する.

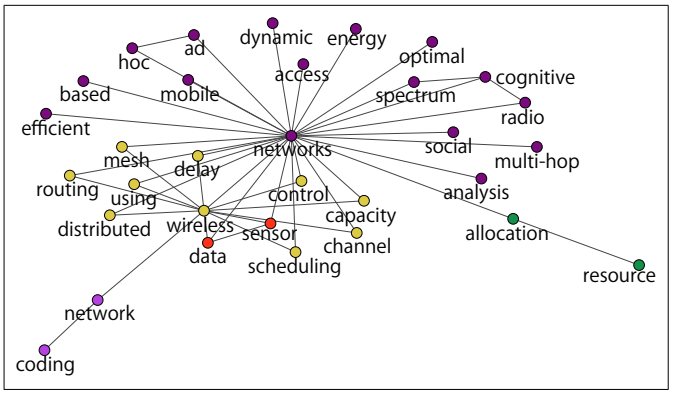
5.1　国際会議 AAAI のトピック発展分析

はじめに，AAAI の論文データセットを $T_{\mathrm{AAAI}} = 5$ 個の時間区間に分割し，提案手法を適用した. なお各時間区間は論文本数がほぼ等しくなるよう設定した. 時間区間 1998-2005, 2006-2008, 2010-2013, 2014-2016 における AAAI の研究トレンド可視化結果を図 3 に示す. ここでは見やすさのため，各時間区間のネットワークのエッジ数が 30 前後となるようパラメータ λ の値を調節した. 図 3 に示すように，提案手法は各時間区間における AAAI の特徴を効果的に抽出できている. 一例として，“learning” という単語周辺の関係は，1998-2005 には “computer” や “language” などの単語と強く関係するのに対し，近年の AAAI では “deep neural networks” や “dictionary learning” の研究が顕著であることがわかる. また図 3 は “semantic web” や “services” のように 2006-2008 の期間を特徴付ける単語のコミュニティも検出されている. 加えて，2010-2013 の期間には，“linked data” や “matrix factorization” のバーストが検出されている. なお図 3 (d) で “what's hot” という語句が検出されているが，これは AAAI2014 から開始されたセッション名に対応しており，AAAI2015, 2016 の予稿集で多くのタイトルがこの語句を含んだことが原因である. このように提案手法は研究トレンドのみならず対象会議の動向や企画も検出できる. 提案手法によって構築した時系列トレンドネットワークは，分野の新規参入者がトピック発展を理解する手助けとなりうる.

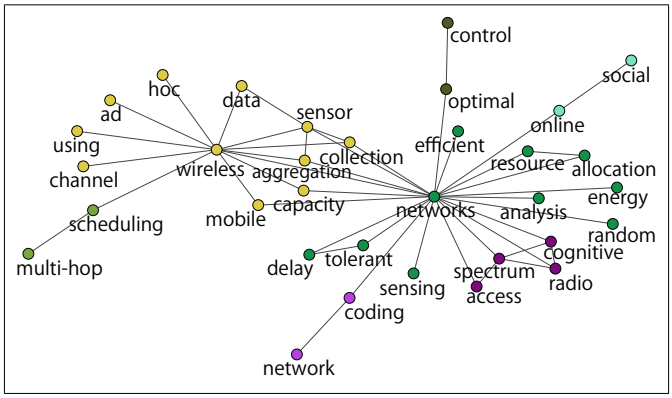
5.2　ベースライン手法との比較

次に，INFOCOM と CVPR を例にとり，提案手法により構築した研究トレンドネットワークと，元の共起語ネットワークを定性的に比較する. 後者は従来のサイエンスマッピングの成果物

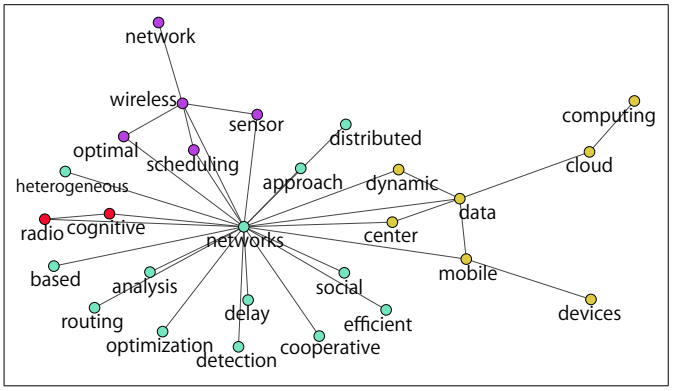
(注6)：我々の予備実験では，提案したスパース－スムース行列分解が L の値に敏感でないことを確認した.



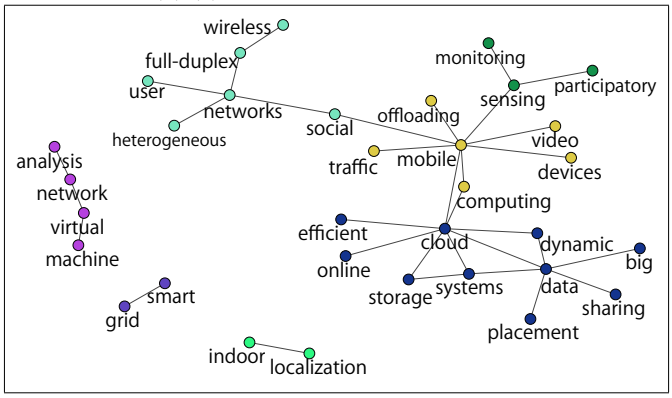
(a) INFOCOM 2010-2012 の共起語ネットワーク（バースト検出前）.



(b) (a) のバースト検出後（提案手法）.

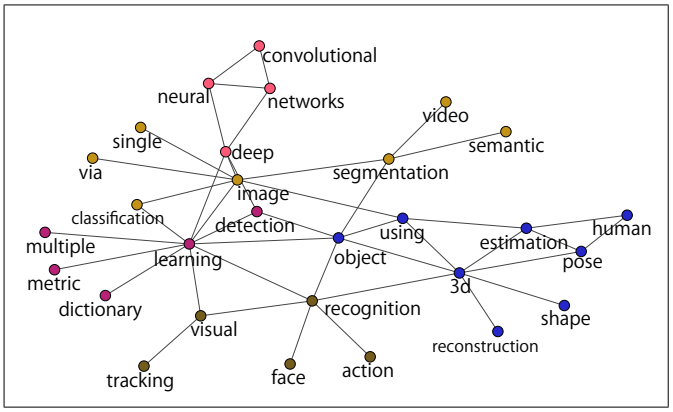


(c) INFOCOM 2013-2016 の共起語ネットワーク（バースト検出前）.

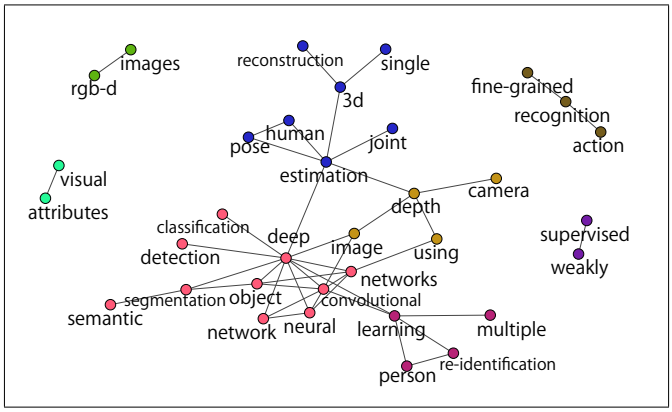


(d) (c) のバースト検出結果（提案手法）.

図 4　INFOCOM 2010-2012 および 2013-2016 におけるバースト検出前後の結果．(a)(b) 元の共起語ネットワーク，(b)(d) 提案手法による研究トレンド可視化結果.



(a) CVPR 2013-2015 の共起語ネットワーク（バースト検出前）



(b) (a) のバースト検出結果（提案手法）.

図 5　CVPR 2013-2015 におけるバースト検出前後の結果．(a) 元の共起語ネットワーク，(b) 提案手法による研究トレンド可視化結果.

に相当する．会議 $c \in \{\text{INFOCOM, CVPR}\}$，$t \in \{1, 2, \cdots, T_c\}$ およびパラメータ λ に対し，$|E(c,t)| = |\mathcal{E}(c,t,\lambda)|$ となるようエッジ数を閾値処理した．各会議 c の論文集合を，ほぼ等しい本数の論文を含む $T_c = 5$ 個の時間区間に分割し，提案手法を適用した．INFOCOM 2010-2012 および 2013-2016 における共起語ネットワークと提案手法によるトレンド検出結果を図 4 に示す．図 4 (a), (c) の元の共起語ネットワークを見比べても，国際会議のトピック発展は理解しにくい．一方，図 4 (b), (d) のトレンドネットワークからは，それぞれの時間区間を特徴付けるクラスタを発見しやすい．例として，図 4 (a), (c) は “networks” を中心に “analysis” や “social” のような定常的な語がつながっている．一方，図 4 (b), (d) では，“full-duplex” や “heterogeneous” などの語が優先されている．同様の傾向は図 5 に示す CVPR の結果からも読み取れる．特に，2013-2015 の CVPR では “RGB-D images” や “fine-grained recognition” のような技術が急激に注目を集めたことがわかる．ゆえに，提案手法は従来のサイエンスマッピングに比べて萌芽的技術が検出でき，それらの可視化はトピック発展の理解に有意義といえる．

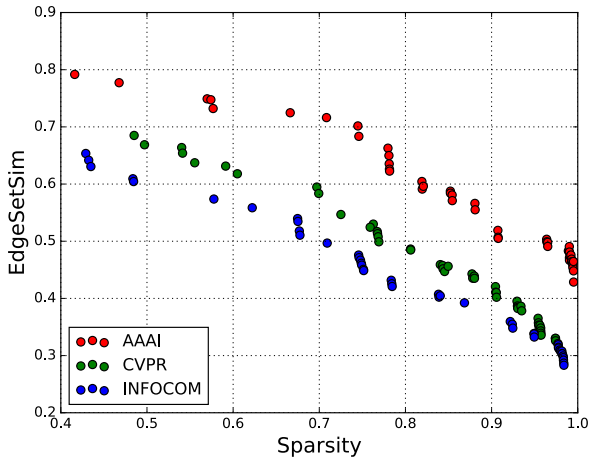


図 6　元の共起語ネットワークとの類似度とスパース性の散布図．縦軸に対して低い値をもつほど，元の共起語ネットワークとは異なるネットワークを抽出したことを表す．

5.3　国際会議の発展スピード分析への応用

提案手法により生成したトレンドネットワークが，元の共起語ネットワークの閾値処理とは異なるエッジ集合を示す場合，元の時系列ネットワークには定常的な語の共起が多く含まれていたと考えられる．つまり，語の共起が定常的であるほど，直前の時間区間から研究内容の変化量が小さいと仮定できる．この仮定に基づき，本節では提案手法から国際会議の発展スピードを表す指標を導出する．まず各会議 c に対し，T_c を開催年数に設定して隣接行列 $\mathbf{W}_c$ を算出する．次に，パラメータ λ で提案手法を行列 $\mathbf{W}_c$ に適用し，スパース行列 $\mathbf{S}_{c,\lambda}$ を得る．得られた行列に基づき，提案手法のもたらしたスパース性を次式により算出する．

$$Sparsity(c,\lambda) = 1 - \frac{\#\text{non-zero entries in } \mathbf{S}_{c,\lambda}}{\#\text{non-zero entries in } \mathbf{W}_c}, \quad (7)$$

$Sparsity(c,\lambda)$ の値が高いほど，元の共起語ネットワークが提案手法によりスパース化されたことを意味する．

加えて，元の共起語ネットワークと提案手法が生成したトレンドネットワークとのエッジ集合類似度を Jaccard 係数に基づき算出する．

$$EdgeSetSim(c,\lambda) = \frac{1}{T_c}\sum_{t=1}^{T_c}\frac{|\mathcal{E}(t,c,\lambda)\cap E(t,c)|}{|\mathcal{E}(t,c,\lambda)\cup E(t,c)|}, \quad (8)$$

$EdgeSetSim(c,\lambda)$ の値が大きいほど，二つのネットワークのエッジ集合は類似していることを表す．つまり，元の共起語ネットワーク自体が十分トレンドを表現するといえる．

パラメータ λ を値域 $[10^{-3}, 10^{-1}]$ で変更し，各国際会議に対するスパース性とエッジ集合類似度の関係をプロットしたものを図 6 に示す．図において，縦軸に対して低い値をもつほど，元の共起語ネットワークとは異なるネットワークを抽出したことを意味する．すなわち，対象とする会議は多くの語の共起が定常的なため，従来のサイエンスマッピングでは流行が捉えにくい．このような研究分野のトレンドの可視化には提案手法が特に有効であるといえる．本実験では，AAAI が最も研究トピックの推移が激しく，INFOCOM が緩やかな推移をもつことが示唆された．今後は実験対象の国際会議数を増やし，さらなる実験と調査を行う予定である．

6. ま と め

本文では，萌芽的技術の変遷に着目した新たなサイエンスマッピング手法として，語の共起のバースト検出に基づく研究トレンドの可視化を提案した．具体的には，時系列共起語ネットワークから急激な時間変化を取り出すためのスパース－スムース行列分解を新たに定式化した．提案手法は単一のパラメータのみでトレンドネットワークのスパース性を操作でき，利便性が高い．また，アルゴリズムの計算コストは行列の要素数に対して線形オーダーであるため，語彙が大規模なデータセットでも高速な計算が可能である．本文の実験では，人工データを用いた定量評価と実データを用いた定性評価により提案手法の有効性を確認した．今後は，実験の規模を拡大することで，提案手法のさらなる評価と応用先を検討する予定である．

本論文は共起語ネットワークの時間発展に着目したが，その他のネットワークからも有用なトレンド情報が検出できる可能性がある．例えば，第一著者はこれまでに研究内容の類似関係 [27,28] や同一組織内の共同研究関係 [29,30] の抽出・分析方法を提案した．今後はこれらの各種ネットワークを入力とした実験も行う．また，エッジ重みのみならず，ノード重みやネットワーク構造の変化も考慮したバースト検出手法を検討する予定である．

文　献

[1] M. Callon, J.-P. Courtial, W. A. Turner, and S. Bauin. From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, Vol. 22, No. 2, pp. 191–235, 1983.

[2] S. G. Assefa and A. Rorissa. A bibliometric mapping of the structure of STEM education using co-word analysis. *Journal of the American Society for Information Science and Technology*, Vol. 64, No. 12, pp. 2513–2536, 2013.

[3] F. Muñoz-Leiva, M. I. Viedma-del Jesús, J. Sánchez-Fernández, and A. G. López-Herrera. An application of co-word analysis and bibliometric maps for detecting the most highlighting themes in the consumer behaviour research from a longitudinal perspective. *Quality & Quantity*, Vol. 46, No. 4, pp. 1077–1095, 2012.

[4] G. A. Ronda-Pupo and L. Á. Guerras-Martin. Dynamics of the evolution of the strategy concept 19622008: a co-word analysis. *Strategic Management Journal*, Vol. 33, No. 2, pp. 162–188, 2012.

[5] J. Zhang, J. Xie, W. Hou, X. Tu, J. Xu, F. Song, Z. Wang, and Z. Lu. Mapping the knowledge structure of research on patient adherence: Knowledge domain visualization based co-word analysis and social network analysis. *PLoS ONE*, Vol. 7, pp. 1–7, 04 2012.

[6] C.-P. Hu, J.-M. Hu, S.-L. Deng, and Y. Liu. A co-word analysis of library and information science in China. *Scientometrics*, Vol. 97, No. 2, pp. 369–382, 2013.

[7] L. Li, X. Li, C. Cheng, C. Chen, G. Ke, D. D. Zeng, and W. T. Scherer. Research collaboration and ITS topic evolution: 10 years at T-ITS. *IEEE Trans. Intelligent Trans-*

portation Systems, Vol. 11, No. 3, pp. 517–523, Sept 2010.

[8] M. Topalli and S. Ivanaj. Mapping the evolution of the impact of economic transition on Central and Eastern European enterprises: A co-word analysis. *Journal of World Business*, 2016.

[9] S. Ravikumar, Ashutosh Agrahari, and S. N. Singh. Mapping the intellectual structure of scientometrics: a co-word analysis of the journal scientometrics (2005–2010). *Scientometrics*, Vol. 102, No. 1, pp. 929–955, 2015.

[10] Y. Liu, J. Goncalves, D. Ferreira, B. Xiao, S. Hosio, and V. Kostakos. CHI 1994-2013: Mapping two decades of intellectual progress through co-word analysis. In *Proc. ACM Conf. Human Factors in Computing Systems (CHI)*, pp. 3553–3562, 2014.

[11] X. Wan, Q. Cheng, and W. Lu. Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks. *Scientometrics*, Vol. 101, No. 2, pp. 1253–1271, 2014.

[12] M. Song, G. E. Heo, and S. Y. Kim. Analyzing topic evolution in bioinformatics: investigation of dynamics of the field with conference data in DBLP. *Scientometrics*, Vol. 101, No. 1, pp. 397–428, 2014.

[13] K. Börner, L. Dall'Asta, W. Ke, and A. Vespignani. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, Vol. 10, No. 4, pp. 57–67, 2005.

[14] N. Shibata, Y. Kajikawa, Y. Takeda, and K. Matsushima. Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, Vol. 28, No. 11, pp. 758 – 775, 2008.

[15] A. E. Motter, A. P. S. de Moura, Y.-C. Lai, and P. Dasgupta. Topology of the conceptual network of language. *Physical Review E*, Vol. 65, p. 065102, Jun 2002.

[16] P. Drieger. Semantic network analysis as a method for visual text analytics. *Procedia - Social and Behavioral Sciences*, Vol. 79, pp. 4 – 17, 2013.

[17] M. L. Doerfel and G. A. Barnett. A semantic network analysis of the international communication association. *Human Communication Research*, Vol. 25, No. 4, pp. 589–603, 1999.

[18] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proc. Int. Conf. Machine learning (ICML)*, pp. 113–120, 2006.

[19] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In *Proc. ACM SIGMOD Int. Conf. Management of Data (SIGMOD)*, pp. 131–142, 2004.

[20] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, Vol. 7, No. 4, pp. 373–397, 2003.

[21] C. Chen. Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, Vol. 57, No. 3, pp. 359–377, 2006.

[22] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, Vol. 2, pp. 183–202, 2009.

[23] J. J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. Comptes Rendus de l ' Académie des Sciences de Paris, Vol. 255, pp. 2897–2899, 1962.

[24] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, Vol. 58, No. 3, pp. 11:1–11:37, June 2011.

[25] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2008, No. 10, p. P10008, 2008.

[26] M. Katsurai. Bursty research topic detection from scholarly data using dynamic co-word networks: A preliminary investigation,. In *Proc. IEEE Int. Conf. Big Data Analysis (ICBDA)*, 2017.

[27] 桂井麻里衣, 大向一輝, 武田英明. 大規模学術論文データベースにおける研究者のトピック推定と著者同定への応用. In *DEIM Forum 2015*, pp. A5–2, 2015.

[28] M. Katsurai, I. Ohmukai, and H. Takeda. Topic representation of researcher's interests in a large-scale academic database and its application to author disambiguation. *IEICE Trans. Information and Systems*, Vol. E99-D, No. 4, pp. 1010–1018, April 2016.

[29] 荒木将貴, 桂井麻里衣, 大向一輝, 武田英明. 研究成果データベースを用いた異分野の共同研究者の推薦. In *DEIM Forum 2016*, pp. E1–3, 2016.

[30] M. Araki, M. Katsurai, I. Ohmukai, and H. Takeda. Interdisciplinary collaborator recommendation based on research content similarity. *IEICE Trans. Information and Systems*, Vol. E99-D, No. 4, April 2017.