DBSJ Japanese Journal
Vol. 16-J, Article No. 1, March 2018

時系列文書に対するトピックフォレストの構築と構造解析

Construction and Structural Analysis of Topic Forest for Time Series Documents

伏見 卓恭♡ 佐藤 哲司♢

Takayasu FUSHIMI Tetsuji SATOH

Web 上には，ニュース記事やブログ記事，ウェブページ，学術文献など，時々刻々と大量の文書が投稿される．文書間には，関連する，あるいは，類似するという関係が強いものと弱いものが存在する．学術文献は引用，ブログ記事はトラックバック，ウィキペディア記事やウェブページはハイパーリンクという形で関連する文書とのつながりが明示されているが，ニュース記事に関しては関連する文書とのつながりが明示されない場合が多い．最も簡単な方法として，ニュース記事間の類似度を計算し，類似文書間にリンクを張ることで類似度ネットワークを構築する方法があげられるが，時間軸を考慮するのは困難である．そこで本稿では，文書の意味的凝集性と時間的凝集性に基づき，時間発展する複数の木構造からなるトピックフォレスト構築手法を提案する．このトピックフォレストを可視化することで，文書への効果的なアクセス順序を提示できると考えられる．実データを用いた評価実験により，構築したトピックフォレストが意味的凝集性と時間的凝集性を有していることを示し，文書群へのアクセシビリティ向上の一助となりうることを確認する．

A large amount of documents are posted on the Web from moment to moment such as news articles, blog articles, web pages, academic literature. There are strong and weak relationships between related and similar documents. The relationships between strongly relevant documents clearly exhibit like citations of scientific literature, trackbacks of blog posts, hyperlinks of Wikipedia articles and web pages, but in the case of news articles, connections with related documents are often not clearly indicated. As a simplest method, there is a method of calculating similarity between news articles and constructing a similarity network by linking between similar documents, but it is difficult to consider the time axis. Therefore, in this paper, we propose a topic forest construction method consisting of multiple time-evolving tree structures based on semantic cohesiveness and temporal cohesion of documents. By visualizing this topic forest, it is considered that an effective access order to the document can be presented. Experimental evaluations using real data show that the topic forest has semantic and temporal cohesiveness, which helps us to improve accessibility to the documents.

♡ 正会員　東京工科大学コンピュータサイエンス学部
takayasu.fushimi@gmail.com

♢ 正会員　筑波大学図書館情報メディア系
satoh@ce.slis.tsukuba.ac.jp

1. はじめに

近年 Web 上には，ニュース記事やブログ記事，学術文献など，大量の文書が時々刻々と投稿されている．文書間には，関連する，あるいは，類似するという関係が強いものと弱いものが存在する．学術文献は引用，ブログ記事はトラックバック，ウィキペディア記事やウェブページはハイパーリンクという形で関連する文書とのつながりが明示されているが，ニュース記事に関しては関連する文書とのつながりが明示されない場合が多い．最も簡単な方法として，ニュース記事間の類似度を計算し，類似文書間にリンクを張ることで類似度ネットワークを構築する方法があげられる．一般的な類似度ネットワーク構築法として，Minimum Spanning Tree，Relative Neighborhood Graph，k-NN Graph などが存在する．これらのグラフは，大規模なオブジェクト群から類似のオブジェクトを検索するタスクである類似検索などでも用いられているが，時間発展する文書群の時間情報を加味するのは困難である．

文書群の全体像を客観的に把握するための手段として，文書間の類似性に基づき，2 次元平面あるいは 3 次元空間に埋め込む可視化手法があげられる [1]．類似性に基づく古典的な可視化法として，Principal Component Analysis，Multi-Dimensional Scaling などが存在する．線形に射影するこれらの可視化手法では高次元ベクトルで表現される文書群を低次元へ埋め込むのが困難であることが知られており，Sammon Mapping や Kernel PCA などの非線形手法も提案されている．さらに，文書のトピック（クラス情報）に着目した可視化手法である Fisher's linear Discriminant Analysis，t-distributed Stochastic Neighbor Embedding [2] や Parametric Embedding [3] も提案されている．これらの可視化手法でも時間情報を考慮するのは困難である．

上述の類似度ネットワーク構築法とネットワーク構造可視化手法を組み合わせることで，高次元多様体上のオブジェクト群を可視化する手法として，ISOMAP [4]，Locally Linear Embedding [5]，Laplacian Eigenmaps [6] があげられる．これらの手法でも，時間情報を付加することは困難である．

本稿では，ISOMAP などのグラフ構造化とグラフ可視化を組み合わせるフレームワークに基づき，時系列文書を効果的に可視化する手法を提案する．時系列文書を適切に可視化することで，文書間の局所的な関係や全体像を客観的に把握できるように整理，かつ所望の文書へのアクセシビリティを高めることを目標とする．具体的には，文書の意味的類似性と時間的凝集性に基づき，時間発展する複数の木構造からなるトピックフォレスト構築手法を提案する．そして，構築したトピックフォレストを可視化する．

本稿は以下のような流れである．2 章で提案手法について述べ，3 章で実データを用いて提案手法を評価し，結果に関して考察する．4 章で関連研究について触れ，最後に本稿のまとめと今後の課題について言及する．

2. 提案手法

提案手法の枠組みでは，文書集合を $\mathcal{D} = \{d_1, \ldots, d_N\}$，単語集合を $\mathcal{W} = \{w_1, \ldots, w_M\}$ とし，各文書は語彙数 M 次元の単語頻度ベクトル $\mathbf{b}_i = [b_{i,j}]_{j=1}^{M}$ で表現する．ここで，$b_{i,j}$ は，文書 d_i における単語 w_j の出現頻度を表す．任意の文書 d_i と d_j 間に類似度 $\rho(d_i, d_j) = \cos(\mathbf{b}_i, \mathbf{b}_j)$ が得られたとき，閾値 α を定め，$\rho(d_i, d_j) > \alpha$ となる文書間にリンクを付与することで，トピックフォレストとよぶ木構造を構築する．構築したトピックフォレストを，角度と半径により座標が決定される極座標平面に埋め込む．全文書集合 $\mathcal{D}$ とパラメータ α を入力として与え，以下の手順で各文書を可視化する：

1. トピックフォレスト構築；

2. 各ツリーを極座標可視化；

各手順の詳細を次節以降で説明する．

2.1 トピックフォレスト構築法

与えられた文書集合に含まれる文書 d をノードとみなし，類似度の高い文書間にリンクを付与することで，ツリー $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ を構築する．具体的には，文書 $d_i \in \mathcal{D}$ が投稿された時刻を $d_i.time$ としたとき，投稿された時刻が早い文書から順にツリーにノードとして追加する．つまり，時間発展するツリーを構築することになる．具体的には，文書 d_i が投稿された時刻より前に投稿された文書集合を $\mathcal{D}^{(d_i)} = \{d \in \mathcal{D}\ ;\ d.time < d_i.time\}$ と表す．文書 d_i について，各文書 $d \in \mathcal{D}^{(d_i)}$ との類似度 $\rho(d, d_i)$ を計算し，最も類似する文書ノード $\hat{d}$ から d_i にリンクを付与する：

$$P(d_i) = \hat{d} = \arg\max_{d \in \mathcal{D}^{(d_i)}} \rho(d, d_i)$$

$$\mathcal{V} \leftarrow \mathcal{V} \cup \{d_i\},\ \ \mathcal{E} \leftarrow \mathcal{E} \cup \{(\hat{d} \longrightarrow d_i)\}$$

ここで $P(d_i)$ は，文書ノード d_i の親ノードを意味する．すなわち，既に投稿されている（＝既にツリーの一部になっている）ノード $d \in \mathcal{D}^{(d_i)}$ のうち，最も類似するノードの子ノードとして，d_i をツリーに追加する．

次に，類似度閾値パラメータ α を導入する．すなわち，文書 d_i について，最大類似度 $\rho(\hat{d}, d_i) = \max_{d \in \mathcal{D}^{(d_i)}} \rho(d, d_i)$ が閾値 α を超える $(\rho(\hat{d}, d_i) > \alpha)$ 場合のみ $\hat{d}$ から d_i にリンクを張り，そうでない場合にはリンクを付与せず，d_i は新たなツリーの根（root）となる．このことを便宜上，$P(d_i) = null$ と記す．

$$P(d_i) = \begin{cases} \arg\max_{d \in \mathcal{D}^{(d_i)}} \rho(d, d_i) & (\rho(\hat{d}, d_i) > \alpha) \\ d_i & (\text{otherwise}) \end{cases}$$

適切な閾値 α を設定することで，異なるトピックの文書群は異なるツリーを形成する．以降，この一連の手順により得られるツリー群をトピックフォレストと呼ぶ．$\alpha < 1$ の値が大きいほどツリーの数は増え，$\alpha = 0$ で単一のツリーとなる．

次に，投稿間隔パラメータ λ を導入する．文書 d_i について，投稿間隔 $d_i.time - d_j.time$ に従いリンク付与確率 $\lambda \exp(-\lambda(d_i.time - d_j.time))$ を与える．これにより，上記の方法で親ノードとして選定した文書ノードとの投稿間隔が大きくなるにつれて，リンク付与確率が低くなる．

各ツリーにおける根 $root$ を 0 階層とし，根からのグラフ距離により各ノードを第 s 階層ノード群 $\mathcal{V}_s = \{d; dist(d, root(d)) = s\}$ のように階層に分ける．第 s 階層のノード数を $N_s = |\mathcal{V}_s|$ とする．また，文書ノード d_i の子ノード集合を $C(d_i)$ とし，その要素を $C(d_i) \in C(d_i)$ と表記する．トピックフォレストにおいて，定義より必ず $P(d_i).time \leq d_i.time$ が成り立つが，$P(d_i)$ の d_i 以外の子ノード d_j に対して，$C(d_j).time > d_i.time$ は成り立つとは限らない．

2.2 極座標可視化法

二部グラフにおける 2 つのノード集合を同心円上に配置する手法 [7] をベースとして，構築したトピックフォレストのそれぞれのツリーを極座標平面上に可視化する Polar Coordinate Embedding 法（以下，PCE 法）について説明する．ツリーにおける第 s 階層と第 $s+1$ 階層ノード群を二部グラフとしてとらえ，座標ベクトル群 $\mathbf{X}_s$ と $\mathbf{X}_{s+1}$ を決定する．$\mathcal{V}_s$ と $\mathcal{V}_{s+1}$ のノードはそれぞれ半径 r_s，r_{s+1} の円上に配置する．

ここでは，第 1 階層と第 2 階層のノード群の座標を決定する方法について説明する．隣接行列 $\mathbf{A} = [a_{i,j}]_{i=1,j=1}^{N_1,N_2}$ に対して多次元尺度法などと同様に中心化を施し，中心化隣接行列 $\tilde{\mathbf{A}} = [\tilde{a}_{i,j}]$ を得る．そして，座標ベクトル群 $\mathbf{X}_1$ と $\mathbf{X}_2$ を適切に初期化し，以下に示す目的関数を最大にするように反復的に求める．

$$J(\mathbf{X}_1, \mathbf{X}_2) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} \tilde{a}_{i,j} \frac{\mathbf{x}_i^T}{r_1}\frac{\mathbf{x}_j}{r_2} + \frac{1}{2}\sum_{i=1}^{N_1}\lambda_i(r_1^2 - \mathbf{x}_i^T\mathbf{x}_i) + \frac{1}{2}\sum_{j=1}^{N_2}\mu_j(r_2^2 - \mathbf{x}_j^T\mathbf{x}_j). \quad (1)$$

ここで，λ_i と μ_j は，各円周上に配置するための制約を表すラグランジュ乗数である．式 (1) において $\frac{\mathbf{x}_i^T}{r_1}\frac{\mathbf{x}_j}{r_2} = \cos\theta_{i,j}$ であり，隣接するノードどうしが原点から見て同じ方向に配置されることによって，$J(\mathbf{X}_1, \mathbf{X}_2)$ は最大化される．すなわち，同じようなノードと隣接するノードどうしを同一方向に，異なるノードと隣接するノードを異なる方向に配置する．

また，ベクトル群 $\mathbf{X}_2$ を固定すれば，第 1 階層ノードの座標ベクトル $\mathbf{x}_i$ の最適配置は以下のように求まる．

$$\mathbf{x}_i = \frac{r_1}{\|\tilde{\mathbf{x}}_i\|}\tilde{\mathbf{x}}_i,\ \ \tilde{\mathbf{x}}_i = \sum_{j=1}^{N_2}\tilde{a}_{i,j}\mathbf{x}_j \quad (2)$$

第 1 階層の文書ノード d_i の座標ベクトル $\tilde{\mathbf{x}}_i$ は，第 2 階層ノードの隣接するノードの座標ベクトル $\mathbf{x}_j$ の合成ベクトルで計算される．そして半径 r_1 上にくるように正規化している．

同様に，ベクトル群 $\mathbf{X}_1$ を固定すれば，第 2 階層ノードの座標ベクトル $\mathbf{x}_j$ の最適配置は以下のように求まる．

$$\mathbf{x}_j = \frac{r_2}{\|\tilde{\mathbf{x}}_j\|}\tilde{\mathbf{x}}_j,\ \ \tilde{\mathbf{x}}_j = \sum_{i=1}^{N_1}\tilde{a}_{i,j}\mathbf{x}_i \quad (3)$$

極座標可視化法のアルゴリズムを以下に示す．

1. ベクトル群 $\mathbf{X}_1$ と $\mathbf{X}_2$ を初期化する；
2. ベクトル群 $\mathbf{X}_2$ を固定し，ベクトル $\mathbf{x}_i$ を求める；
3. ベクトル群 $\mathbf{X}_1$ を固定し，ベクトル $\mathbf{x}_j$ を求める；
4. 目的関数 $J(\mathbf{X}_1, \mathbf{X}_2)$ の変化が十分小さければ終了する；
5. （2）へ戻る；

このアルゴリズムは HITS アルゴリズム [8] と類似した構造を持つことが分かる．ただし，ベクトル群に対して 2 重の中心化を施す点，および，正規化の施し方の点に特徴を持つ．提案アルゴリズムの 1 反復は，2 部グラフのリンク数に比例した計算量となる．よって，ネットワーク可視化の代表手法の一つバネモデル法 [9] などの非線形最適化が必要な可視化法と比較して，高速な方法である．

上記の手順により，第 1 階層と第 2 階層ノードの最適配置が求まった．次いで，第 2 階層と第 3 階層ノードの関係性からそれらの最適配置を求める．

3. 評価実験

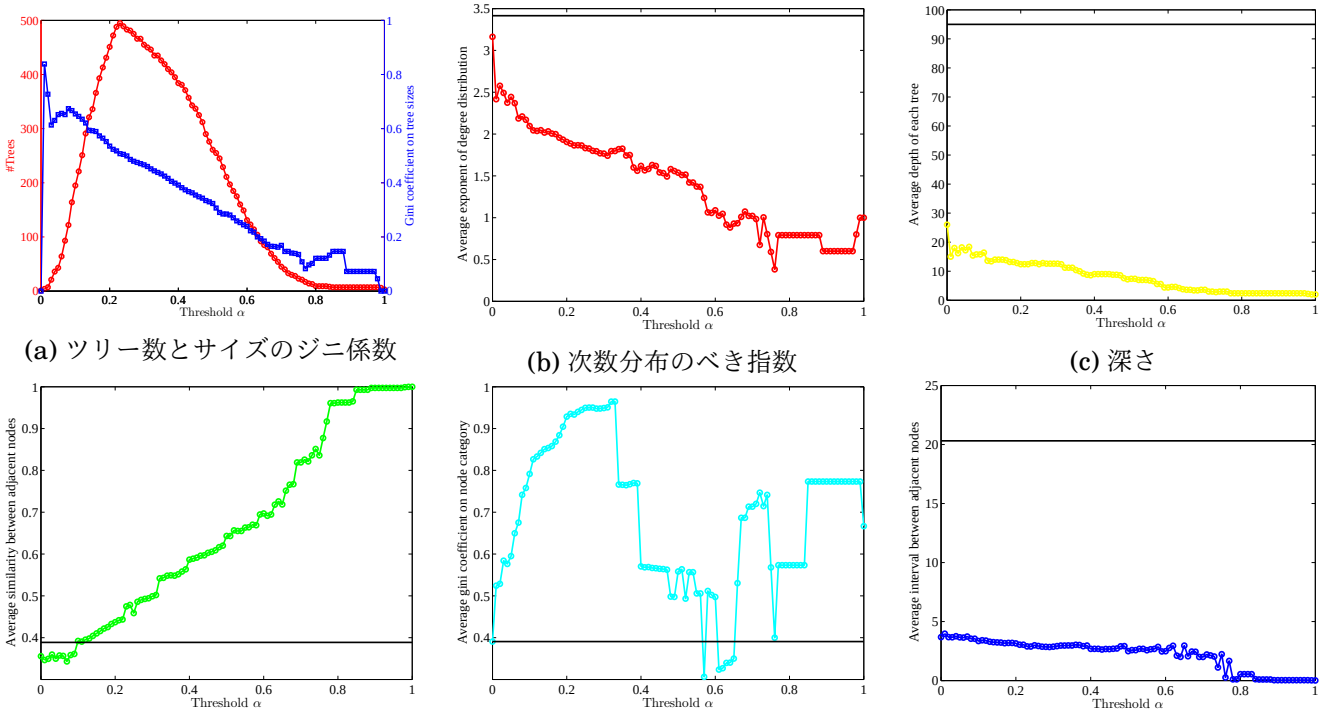
実ニュース記事データに対する提案手法の処理結果について，1) 構築したトピックフォレスト（TF）の構造，2) 高次数ノード，3) 可視化，4) アノテーションの観点から Minimum Spanning Tree（MST）と比較して評価する．

3.1 データセット

提案手法の評価に際して，2014 年 4 月から 6 月までの Yahoo! ニュースの 6,450 記事を用いる．含まれる語彙数は 31,090 であり，記事間の類似度としてベクトル空間モデル [10] で頻繁に用いられる単語頻度ベクトル（Bag Of Words）間のコサイン類似度を採用する．データ中の主要な記事として，小保方晴子氏の STAP 細胞問題，田中将大投手の MLB 移籍，セウォル号沈没事故などが含まれる．

3.2 トピックフォレストの構造

提案手法により構築したトピックフォレスト（TF）について，閾値パラメータ α を変化させた際の構造について評価する．図 1 に，横軸を α，縦軸を各種統計量の値としたグラフを示す．図中の黒い実線は，MST における統計量の値を示す．MST は，全文書ノードを対象に，類似度 ρ にしたがって構築する．



(a) ツリー数とサイズのジニ係数　(b) 次数分布のべき指数　(c) 深さ

(d) 平均隣接類似度　(e) カテゴリジニ係数　(f) 平均隣接投稿間隔

図 1: 閾値パラメータ α に対する TF 構造の変化

図 1(a) は，TF を構成するツリー数の推移と各ツリーに含まれる文書ノード数のジニ係数の値である．ジニ係数は，各ツリーのノード数に大きな差がある場合に 1 に近づき，均等な場合に 0 に近づく．MST は単一のツリーであり結果は自明であるため，この図には示していない．なお，パラメータ α におけるツリー数を K_α，ツリー $\mathcal{T}^k$ $(k = 1, \ldots, K_\alpha)$ に含まれるノード数を $N^{(k)} = |\mathcal{V}^{(k)}|$ とし，ジニ係数は以下のように計算した：

$$\frac{\sum_{k=1}^{K_\alpha - 1} \sum_{h=1}^{K_\alpha} |N^{(k)} - N^{(h)}|}{(K_\alpha - 1) \sum_{k=1}^{K_\alpha} N^{(k)}}. \tag{4}$$

ただし，ノード数 1 のツリーはカウントしない．図 1(a) からわかることとして，

- $\alpha = 0$ の際には単一のツリーとなる．
- α が大きくなるにつれてジニ係数は減少する．
- $\alpha = 0.23$ でツリー数が最大となる．

ノード数 1 のツリーもカウントすれば，ツリー数は単調増加となることに注意されたい．

図 1(b) に，以下の式に示す次数分布のべき指数 a の値をプロットした．

$$DegDist(degree) \propto \times degree^{-a}$$

図 1(b) から，MST と比較して TF のべき指数は小さいことがわかるが，$0 < \alpha < 0.2$ あたりで $2 < a < 3$ となり，実ネットワークと同様にスケールフリー性を有することがわかる．単一ツリーとなる $\alpha = 0$ においても，TF は MST と幾分か差があることがわかる．MST は，TF の総ノード数，総リンク数と等しいが，次数分布は異なることが明らかとなった．

図 1(c) に，TF の各ツリーの深さの平均と MST の深さプロットした．図 1(c) から，両者の深さには大きな違いがあることが見て取れる．単一ツリーとなる $\alpha = 0$ においても，大きな違いがある．この違いは，後述する可視化結果にも表れており，類似の文書を探索するという観点では，TF の方が優れていると言える．

図 1(d) に，以下に示す平均類似度の値をプロットした：

$$\frac{1}{|\mathcal{V}|} \sum_{d \in \mathcal{V}} \rho(d, P(d))$$

以下，この値を平均隣接類似度と呼ぶ．図 1(d) から，α の値を大きくするにつれて，平均隣接類似度も高くなることがわかる．これは当然の結果である．なぜならば，TF 構築時に，$\rho(d_i, d_j) > \alpha$ である文書間にリンクを付与しているからである．すなわち，図 1(d) においては，平均隣接類似度は必ず閾値 α より大きくなる．さらに，$\alpha > 0.1$ で MST よりも高い値になることがわかる．

図 1(e) に，各文書に付与されているカテゴリの偏り具合をジニ係数により計算した値をプロットした．すなわち，ジニ係数の値が高いほど，各ツリーには限られたカテゴリが偏って含まれることを意味する．なお，カテゴリジニ係数は，式 (4) に準じて計算する．逆にジニ係数の値が低いほど，多くのカテゴリが均等に含まれていることを示す．図 1(e) から，MST と比較して TF の方が有意にジニ係数が高いことがわかる．MST では全ノードが単一のツリーに属するため，MST におけるカテゴリジニ係数は全データのカテゴリの偏り具合を表現しているが，その値より有意に高いことが分かる．閾値を設定して異なるトピックの文書ノードが異なるツリーに属するため，このような結果になり，TF の狙い通りの結果になったと言える．以上のことより，MST では単一のツリーとなり，異なるトピックの文書ノードが含まれるが，TF では異なるトピックの文書ノードは異なるツリーとなり，各ツリーには意味的凝集性があることが示された．

図 1(f) に，以下の式で計算される隣接する文書ノード間の投稿された時刻の間隔の平均値（以下，平均隣接投稿間隔）をプロットした：

$$\frac{1}{|\mathcal{V}|} \sum_{d \in \mathcal{V}} |d.time - P(d).time|.$$

図 1(f) から，MST と比較して隣接ノード間の投稿間隔は小さい，

表 1: $\alpha = 0.23$ の TF における高次数ノードの例

親	【次数 15】マー君 メジャー初登板で勝利
子 1	マー君初勝利 何度もうれしい
子 2	里田まい 特等席でマー君応援
子 3	マートン 7 打点でも阪神連敗
子 4	ダル今季初勝利 エースの底力
子 5	黒田 本拠地開幕戦で初勝利
子 6	ヤ軍 16 失点大敗 内野手が登板
子 7	マー君 ブーイング必至も不敵
子 8	マー君 2 者連続被弾も 3 勝目
子 9	日本人トリオ 依存高まるヤ軍
子 10	田中 4 失点降板 連勝止まるか
子 11	前回黒星マー君 1 失点で 7 勝目
子 12	田中 8 勝 防御率トップを維持
子 13	マー君 月間 MVP に初選出
子 14	第 2 のマー君 米メディア特集
子 15	「14 年版マー君」に川崎脱帽

親	【次数 11】小保方氏 理研で研究続けたい
子 1	Twitter にみる小保方氏会見
子 2	小保方氏 理研に不服申し立て
子 3	小保方氏応援 tweet 批判の 2 倍
子 4	iPS 臨床研究に向け準備着々
子 5	慈恵医大 科研費を不正申請か
子 6	偽医者 県財団だまし講演 14 回
子 7	笹井氏 STAP 論文撤回が適切
子 8	山中教授 過去の切り貼り否定
子 9	降圧剤 千葉大論文も改ざんか
子 10	小保方氏の指導役 STAP は本物
子 11	理研 STAP で遠のく新法人指定

親	【次数 6】ケミストリー堂珍が離婚へ
子 1	堂珍離婚報道 妻サイドは否定
子 2	つちや ふっくんと離婚の決意
子 3	能世あんな、重松隆志と離婚
子 4	A・バンデラス、女優と離婚か
子 5	高岡早紀の妹 離婚報道に驚き
子 6	カイヤ 離婚は「分からない」

すなわち，より関連の強い文書ノード同士がつながっており，各ツリーには時間的凝集性があることが示唆された．

3.3 高次数ノード

次に，TF において高次数なノードを 3 ノード取り上げ，その性質について定性的に評価する．上述の評価実験の結果を参考にして，表 1 にツリー数が最も多くなった $\alpha = 0.23$ における高次数ノードを示す．表 1 を見ると，データ中で初期に投稿された「マー君 メジャー初登板で勝利」という記事は，次数 15 でツリーの中では高次数ノードである．その子ノード群を見ると，同様の内容の「マー君初勝利 何度もうれしい」や田中将大投手の妻である”里田まい”に関する記事「里田まい 特等席でマー君応援」が存在する．これらの記事の子孫として，”田中将大投手”に関する連日の話題が多くを占めていた．また，同一の内容ではないものの関連する”マートン”，”黒田投手”，”ダルビッシュ投手”なども子ノードとして存在している．これらの記事の子孫として，”田中将大投手”の内容ではなく”黒田投手”や”ダルビッシュ投手”などの話題が連なっており，高次数ノード「マー君 メジャー初登板で勝利」を機に分岐していた．

高次数ノード「小保方氏 理研で研究続けたい」においても同様の傾向が見受けられた．”小保方氏本人”に関する記事，所属する組織である”理研”に関する記事，同分野の他の教授”山中教授”や”科学技術分野における不正”に関する記事を子ノードとしてもち，さらにそれを機に分岐している．

高次数ノード「ケミストリー堂珍が離婚へ」においても同様の傾向が見受けられた．離婚というキーワードで他の夫婦（つちやかおり／布川敏和，能世あんな／重松隆志など）の離婚話題が子ノードとしてつながっており，その子孫として各夫婦の話題に分岐している．

このように，高次数ノードは関連する記事を子ノードとして多く持ち，より密に関連する記事群へ枝分かれする分岐点の役割を果たしていることが示唆された．

3.4 可視化結果

次に，提案手法における可視化技術である極座標可視化法（PCE 法）について評価する．以下の代表的なネットワーク可視化手法と比較する．構築したグラフの隣接行列を $\mathbf{A} = [a_{u,v}]$，グラフ上でのノード u, v 間の距離を要素とした行列を $\mathbf{G} = [g_{u,v}]$，全ノードの座標ベクトル群を $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}^T$ とする．

- (b) ばねモデル（Spring-Force 法）

$$\mathcal{K}(\mathbf{X}) = \sum_{m=1}^{N-1} \sum_{n=m+1}^{N} \frac{1}{2g_{m,n}^2} (g_{m,n} - \|\mathbf{x}_m - \mathbf{x}_n\|)^2$$

- (c) クロスエントロピー法（Cross-Entropy 法）

$$C(\mathbf{X}) = -\sum_{m=1}^{N-1} \sum_{n=m+1}^{N} \{a_{m,n} \ln \rho(m,n) + (1-a_{m,n}) \ln(1-\rho(m,n))\}$$

- (d) グラフ距離に基づく多次元尺度構成法（Graph-Multi-Dimensional-Scaling 法）

$$\mathcal{M}(\mathbf{X}) = \frac{1}{2} \sum_{h=1}^{2} \mathbf{x}_{(h)}^T (\mathbf{H}_N \mathbf{G} \mathbf{H}_N) \mathbf{x}_{(h)}$$

- (e) スペクトラル埋込法（Spectral-Embedding 法）

$$\mathcal{S}(\mathbf{X}) = \sum_{h=1}^{2} \mathbf{x}_{(h)}^T \mathbf{L} \mathbf{x}_{(h)}$$

により $\alpha = 0$ とした際の TF を可視化した結果を示す．ここで，$\mathbf{H}_N = \mathbf{I}_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T$ は中心化行列，$\mathbf{L}$ はグラフラプラシアン行列，$\mathbf{x}_{(h)}$ は全ノードの座標ベクトルのうち h 次元の値を並べたベクトルである．また，$\rho(m,n) = \exp(-\|\mathbf{x}_m - \mathbf{x}_n\|^2/2)$ とした．さらに，多様体学習の代表手法である (f) ISOMAP （k-NN 法によるグラフ構築，ダイクストラ法によるノード間距離計算，多次元尺度構成法による可視化）による結果も示す．

図 2 に，各可視化結果を示す．図中のノードの色は，対応する文書のカテゴリを表している．図 2(a) を見ると，時系列文書が投稿時刻の順に従い同心円状に，かつ，リンクで結ばれた関連文書や同一カテゴリの文書（同一色のノード）が原点から見て同一方向に布置されており，半径と角度という軸により興味のある文書へのアクセシビリティが高いように見受けられる．図 2(b) と (c) を見ると，ツリー構造がリンクの重なりなく平面に埋め込めるた

め，効率よく平面上に布置されており，かつ，同一カテゴリ文書が近傍に布置されていることがわかる．しかし，提案手法と比較すると，文書の投稿順序は可視化結果に陽に反映されていない．さらに，SF 法と CE 法はニュートン法に基づく非線形最適化をしているため，布置座標を計算するのに多大な時間がかかる．一方 PCE 法では，HITS アルゴリズムなどと同様なパワー法に基づく最適化をしているため，非常に高速に布置座標を計算できる．

図 2(d) と (e) を見ると，ノードが重なり合い文書間の関係が不鮮明である．G-MDS 法と SE 法は線形な手法であるため，高速である一方で解の品質が良くなく，このような結果になったと考えられる．

図 2(f) を見ると，同一カテゴリの文書が近傍に位置し，文書間の関連が反映された結果となっている．ツリーを構築する提案手法とは異なり k-NN グラフを構築するため閉路が存在し，文書間の順序関係が反映できない．一方提案手法では，TF 構築時に順序関係を考慮しているため，文書群の流れが陽にわかる可視化結果を実現できている．

最後に図 3(b) は，投稿時刻は考慮せず文書間の類似度だけで構築した MST をばねモデルで可視化したものである．前述したとおり，ツリー全体の深さや直径が大きい点が TF との顕著な違いである．

3.5 アノテーション

この節では，提案手法の出力である可視化結果において，原点から見て同一方向に関連する文書ノード群が布置されているかについて，アノテーションの観点から評価する．PCE 法により求めた N 個の文書ノードの H 次元[1]座標ベクトル群 $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}^T$ を用いる．ここで，座標ベクトルは重心が $\mathbf{0}$ に，各座標値の自乗和が 1 となるように正規化されているとする．すなわち，$\sum_{i=1}^{N} \mathbf{x}_i = \mathbf{0}$, 任意の次元 h $(1 \leq h \leq H)$ で $\sum_{i=1}^{N} x_{i,h}^2 = 1$ である．一方，各文書に出現する単語 j の出現頻度を用いて N 次元の属性値ベクトルを $\mathbf{y} = [b_{1,j}, \cdots, b_{N,j}]^T$ とし，H 次元の射影ベクトル $\mathbf{f}$ とする．ここで，$\|\mathbf{f}\| = \sum_{h=1}^{H} f_h^2 = 1$ とする．このとき，ベクトル $\mathbf{f}$ 上への座標ベクトル群の射影値から構成される N 次元縦ベクトルは $\mathbf{Xf}$ となる．よって，属性値ベクトル $\mathbf{y}$ をアノテートする妥当な方向として，次式を最大にする射影ベクトル $\mathbf{f}$ を考える．

$$F(\mathbf{f}) \quad = \quad \mathbf{y}^T \mathbf{X} \mathbf{f}. \tag{5}$$

PCE 法により得られた座標ベクトル群 $\mathbf{X}$ を上述のように正規化することで，式 (5) で定義した $F(\mathbf{f})$ はベクトル $\mathbf{Xf}$ と $\mathbf{y}$ の相関係数と等価になる．よって，式 (5) で定義した $r = F(\mathbf{f})$ を以下では簡単に相関と呼ぶ．

式 (5) の相関を最大化する $\hat{\mathbf{f}}$ はラグランジュ乗数法より以下となる．

$$\hat{\mathbf{f}} \quad = \quad \frac{1}{\|\mathbf{X}^T \mathbf{y}\|} \mathbf{X}^T \mathbf{y}. \tag{6}$$

一方，式 (6) を式 (5) に代入すれば以下を得る．

$$F(\hat{\mathbf{f}}) \quad = \quad \|\mathbf{X}^T \mathbf{y}\|. \tag{7}$$

よって，属性値ベクトル $\mathbf{y}$ を H 次元空間に埋め込むアノテーションとして，その方向と相関を，それぞれ式 (6) と 式 (7) で規定する次式のベクトル（矢印）と定義する．

$$\mathrm{Annot}(\mathbf{y}) \quad = \quad \mathbf{X}^T \mathbf{y}. \tag{8}$$

明らかに，属性値ベクトル $\mathbf{y}$ に対して，式 (6) の矢印が長ければ相関が高く有意なアノテーションと言えるが，矢印が短ければアノテーションが困難なことを意味する．本稿では，全ての単語 j に関して，属性値ベクトル $\mathbf{y}$ を構築しアノテーションを試みる．そして，相関が上位の単語によりアノテーションを付与することで，提案手法による可視化結果において，上位単語で特徴づけられる関連文書群が原点から見て同一方向に密集しているか評価する．

図 4 に，$\alpha = 0.23$ の TF のうち大きなツリー 3 つに対するアノテーション結果を示す．相関ランキング上位 5 単語によるアノテーション矢印を示す．図中矢印の色は，キャプションに示している単語によるアノテーションを意味する．ノードの色は，対応する文書のカテゴリを意味する．図 4(a) [2]では，赤い矢印の方向に「イチロー」という単語が出現する文書ノードが多く存在することを表している．他の矢印でも同様に，各単語が出現する文書ノードが矢印の方向に多く存在することを意味する．図 4(a) の結果より．提示した 5 単語に関して高い相関係数を示しており，同一単語を有する関連文書が原点から見て同一方向に布置されていることが示された．

図 4(b) では，上位 5 単語によるアノテーションの結果がすべて同一方向を指している．右上および右下方向には，「ロシア・ウクライナ情勢」に関する文書が散在しており，方向的に広く分布しているため，布置座標と単語分布に強い相関が得られない．これが起因して，相関上位の単語として，方向的にまとまっている「イラク情勢」に関する文書ばかりが得られてしまったと考えられる．

図 4(c) では，図 4(a) と同程度に高い相関係数が得られ，各矢印の方向に同一単語が出現する関連文書が布置されており，提案手法による可視化結果の有用性が示唆された．

4. 関連研究

本稿では，時系列文書群に対して，投稿順序を保持し，類似する文書群をつなぐことによりツリー群を構築し，効果的に可視化する手法を提案した．時系列文書可視化の関連研究として，Ishikawa らの T-Scroll がある [11]．このシステムでは，時間的に離れた文書間の影響力が，指数的に小さくなる重みを導入した類似度を定義し，k-means 法によりインクリメンタルなクラスタリングを実現している．そして，隣接時刻間のクラスタ間に関連度を定義し，関連度の強いクラスタ間にリンクを付与することで，各時刻におけるクラスタリング結果を時間軸上にプロットしている．本稿のトピックフォレスト構築時にも，投稿間隔と文書間の類似度を考慮する点で関連する研究であるが，各文書を極座標平面に布置する点，クラスタとしてまとめず各文書を可視化する点，ツリー構造としての文書間の関係を表現する点で異なる．

キーワード抽出や特徴選択の技術を用いて重要な単語やトレンドワードを抽出し，その頻度をプロットすることで時系列全体を俯瞰する手法として，MemeTracker [12] や EventRiver [13]，CloudLines [14]，STREAMIT [15] などがある．Keim らの EventRiver [13] では，日々の出来事（イベント）について記されたニュースやブログの記事群に対して，短いタイムスパンのバケツに文書を分ける．バケツ内の文書をクラスタリングすることで，temporal-locality クラスタと呼ぶ時間的凝集性と意味的凝集性の高い文書群に分割する．そして，得られたクラスタ群をメタクラスタリングすることで，時間的に離れていても内容的に類似するような，長期間にわたるイベントに関する文書クラスタを一つのグループとしてまとめあげることができる．横幅をグループの生存時間，縦幅を各時点での影響度とした成長曲線を描くことで，イベントの起点や終点，盛り上がりなどが一目瞭然となり，関連イベントの発見も可能にする可視化結果を実現している．

Krstajic らの CloudLines [14] は，あるキーワードを含むニュース記事群を時間軸上に点としてプロットする．カーネル密度推定により投稿間隔が密なイベントを重要なイベントとし，点のサイズや透明度に反映させる．プロットされた点群は時間軸上で連なることで線となり，重要なイベントが発生した箇所では濃く太い

[1] 本稿では 2 次元極座標平面に可視化しているため，$H = 2$ である．

[2] 少し見づらいが，黄矢印は 2 時 30 分の方向，桃矢印は 5 時の方向を指している．

線となる．このような工夫により，短時間で出現頻度の高い記事を目立たせる可視化手法を提案している．イベントの全体像を俯瞰することができ，各イベントを拡大することでその詳細を把握することができる．

これらの既存技術では，横軸に時間軸を設定し，各トピックやトレンドの変化をプロットすることで全体像を把握できるように工夫されている．さらに，ユーザが興味あるトピックやトレンドを選択することで，より詳細な情報を入手できる点で優れている．本稿の提案手法がこれら既存技術と異なる最大のポイントは，各話題やトピックが時間発展する中で複数の観点に分岐する様子をツリー構造により表現できる点である．各話題においては，一般的に情報が広がっていく傾向にあるため，極座標可視化のように，原点付近では描画領域が限られていても，原点から離れる（半径が大きくなる）につれて描画領域が広くなるのもメリットの一つである．

Alsakran らの STREAMIT [15] は，あるスナップショットにおける文書群を粒子として扱い，文書粒子間の類似度に基づき力学モデルにより 2 次元平面での最適配置を計算する．すなわち，類似する文書粒子は平面上で近傍に，類似しない文書粒子は遠くに配置される．新規に追加される文書に対して，動的に最適配置を再計算し，各スナップショットでの最適配置をアニメーションにより表示する．平面に配置された文書粒子に対して，ドロネー三角形によるグラフを構築し，閾値パラメータより長いリンクを切断することで，文書クラスタを形成する．この点は，閾値より離れた文書ノードを別のツリーとする提案手法と類似している．一方で，極座標平面の半径により時間軸を表現する我々の提案手法とは大きく異なり，アニメーションにより対応する文書クラスタの変化を表現している．すなわち，STREAMIT は，時系列データの俯瞰という観点ではデメリットがある．

5. おわりに

本研究では，ユーザの興味ある文書へのアクセシビリティ向上をめざし，時々刻々と投稿される時系列文書群を意味的凝集性と時間的凝集性の観点で適切に可視化する手法を提案した．提案手法におけるトピックフォレスト（TF）は，類似の文書ノードの投稿順序を保存した構造を実現し，極座標可視化法（PCE）は，原点から見て同一方向に関連文書を布置できることを実験により確認した．

今後の課題として，各文書が投稿された時刻を極座標可視化法における半径に反映するなどの拡張をすることで，より効果的な可視化をめざす．TF では，ツリー構造により話題の発展，分岐を表現できたが，話題の収束に関しては表現できない．複数の文書をまとめる文書を DAG 構造などにより表現することも今後の課題となる．さらに，本稿の実験では，TFIDF のコサイン類似度を用いたが，分散表現などによる文書間の類似度を用いることも理論上は可能であるため，他の類似度指標と可視化法の親和性についても検証が必要である．

[謝辞]

本研究は，JSPS 科研費 16K16154 の助成を受けたものです．

[文献]

[1] A. Šilić and B.D. Bašić, “Visualization of text streams: A survey,” pp.31–43, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[2] L. van derMaaten and G.E. Hinton, “Visualizing high-dimensional data using t-sne,” Journal of Machine Learning Research, vol.9, pp.2579–2605, 2008.

[3] T. Iwata, K. Saito, N. Ueda, S. Stromsten, T.L. Griffiths, and J.B. Tenenbaum, “Parametric embedding for class visualization,” Advances in Neural Information Processing Systems 17, eds. by L.K. Saul, Y. Weiss, and L. Bottou, pp.617–624, MIT Press, 2005.

[4] J.B. Tenenbaum, V. Silva, and J.C. Langford, “A Global Geometric Framework for Nonlinear Dimensionality Reduction,” Science, vol.290, no.5500, pp.2319–2323, 2000.

[5] S.T. Roweis and L.K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” SCIENCE, vol.290, pp.2323–2326, 2000.

[6] M. Belkin and P. Niyogi, “Laplacian eigenmaps and spectral techniques for embedding and clustering,” Advances in Neural Information Processing Systems 14, pp.585–591, MIT Press, 2001.

[7] T. Fushimi, Y. Kubota, K. Saito, M. Kimura, K. Ohara, and H. Motoda, “Ai 2011: Advances in artificial intelligence: 24th australasian joint conference, perth, australia, december 5-8, 2011. proceedings,” chapter Speeding Up Bipartite Graph Visualization Method, pp.697–706, Springer Berlin Heidelberg, 2011.

[8] J.M. Kleinberg, “Authoritative sources in a hyperlinked environment,” J. ACM, vol.46, pp.604–632, Sept. 1999.

[9] T. Kamada and S. Kawai, “An algorithm for drawing general undirected graphs,” Inf. Process. Lett., vol.31, pp.7–15, April 1989.

[10] G. Salton, A. Wong, and C.S. Yang, “A vector space model for automatic indexing,” Commun. ACM, vol.18, no.11, pp.613–620, Nov. 1975.

[11] Y. Ishikawa and M. Hasegawa, “T-scroll: Visualizing trends in a time-series of documents for interactive user exploration,” pp.235–246, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[12] J. Leskovec, L. Backstrom, and J.M. Kleinberg, “Meme-tracking and the dynamics of the news cycle,” KDD, pp.497–506, 2009.

[13] D.A. Keim, D. Luo, J. Yang, W. Ribarsky, and M. Krstajic, “Eventriver: Visually exploring text collections with temporal references,” IEEE Transactions on Visualization & Computer Graphics, vol.18, pp.93–105, 2010.

[14] M. Krstajic, E. Bertini, and D.A. Keim, “Cloudlines: Compact display of event episodes in multiple time-series.,” IEEE Trans. Vis. Comput. Graph., vol.17, no.12, pp.2432–2439, 2011.

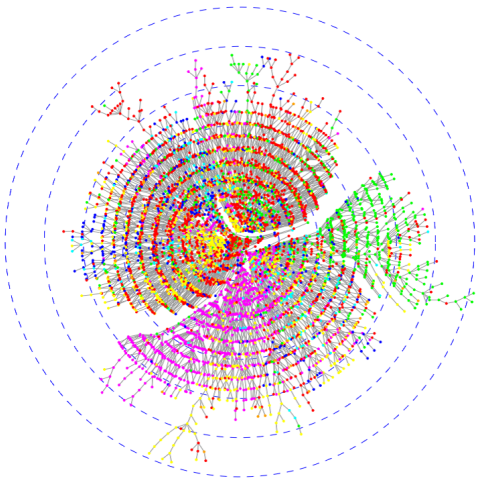
[15] J. Alsakran, Y. Chen, D. Luo, Y. Zhao, J. Yang, W. Dou, and S. Liu, “Real-time visualization of streaming text with a force-based dynamic system,” IEEE Comput. Graph. Appl., vol.32, no.1, pp.34–45, Jan. 2012.

伏見 卓恭 Takayasu FUSHIMI

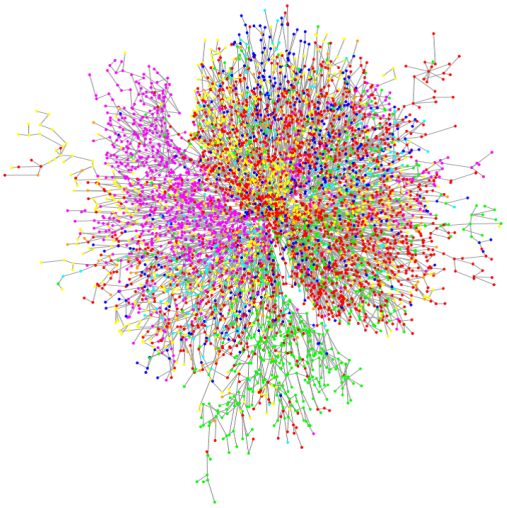
東京工科大学コンピュータサイエンス学部助教．2014 年静岡県立大学大学院経営情報イノベーション研究科博士後期課程修了．同年静岡県立大学大学院経営情報学部客員研究員．2015 年筑波大学図書館情報メディア系特別研究員（PD）．2017 年より現職．複雑ネットワーク解析，可視化の研究に従事．博士（学術）．人工知能学会，情報処理学会各会員．

佐藤 哲司 Tetsuji SATOH

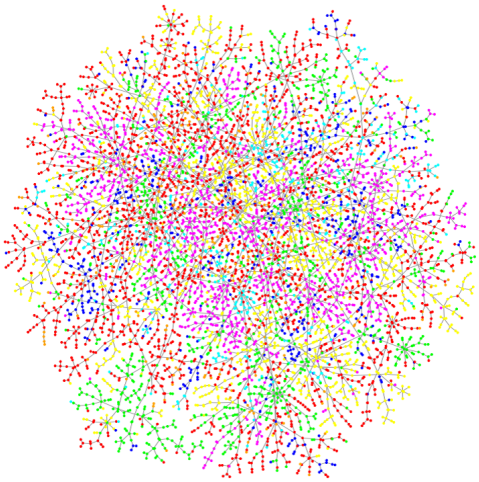
筑波大学図書館情報メディア系教授．1980 年山梨大学工学部電子工学科卒業．同年日本電信電話公社（現 NTT）武蔵野電気通信研究所に入所．以来，データベースマシン，マルチメディアデータベース，情報検索・情報共有の高次化などに関する研究・開発に従事．2007 年 4 月より現職．情報検索・知識発見，社会ネットワーク分析、社会インタラクションに興味を持つ．情報処理学会，電子情報通信学会，日本データベース学会各会員．大阪大学博士 (工学) .



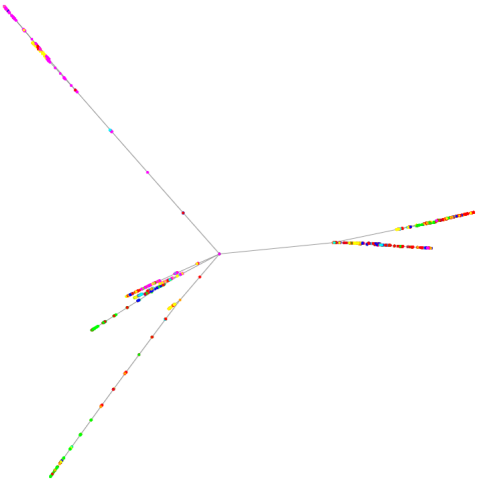
(a) 提案手法（TF+PCE）



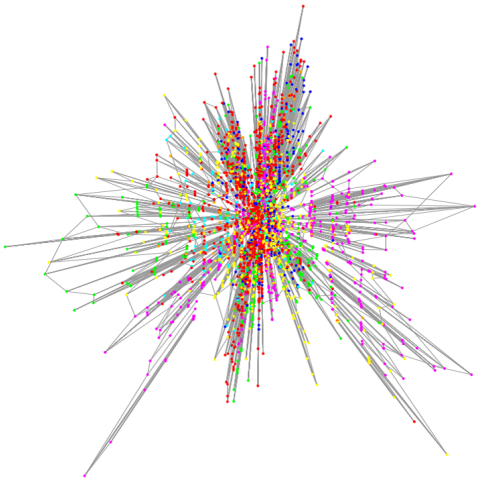
(b) TF+SF



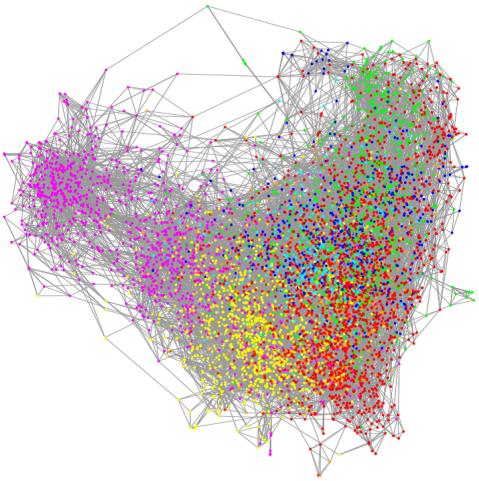
(c) TF+CE



(d) TF+G-MDS

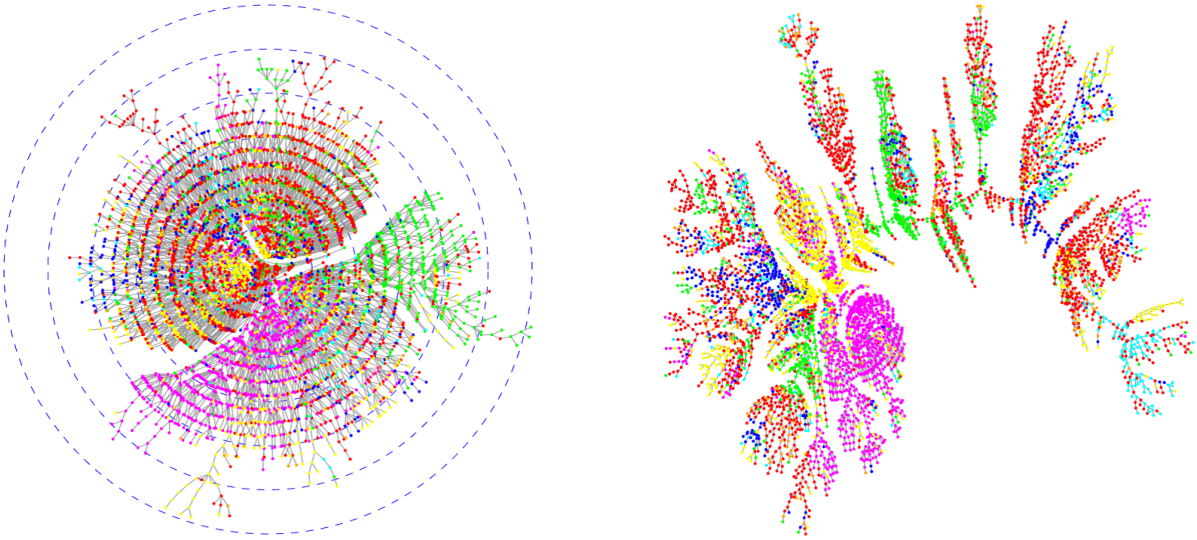


(e) TF+SE



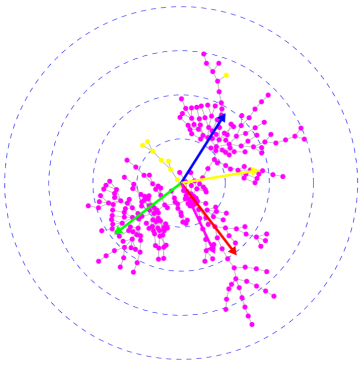
(f) ISOMAP（4-NN+G-MDS）

図 2: $\alpha = 0$ の TF の可視化結果

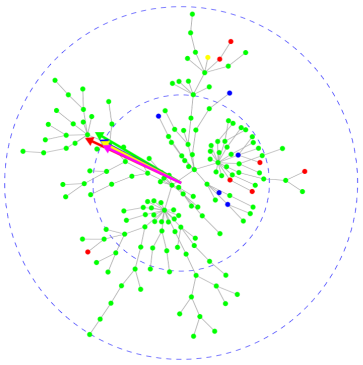


(a) 再掲：提案手法（TF+PCE）　　(b) MST+SF

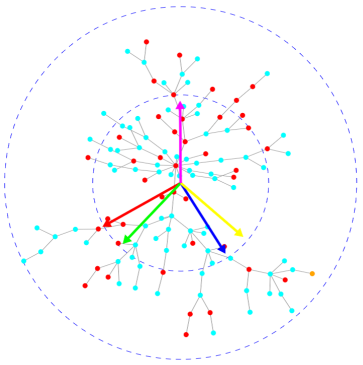
図 3: $\alpha = 0$ の TF の可視化結果



(a) 第 1 ツリー （野球）
赤矢印：「イチロー」$r = 0.49$,
緑矢印：「西武」$r = 0.45$,
青矢印：「田中」$r = 0.44$,
黄矢印：「ダルビッシュ」$r = 0.42$,
桃矢印：「外野」$r = 0.41$
桃ノード：カテゴリがスポーツの文書



(b) 第 2 ツリー （中東情勢）
赤矢印：「イラク」$r = 0.57$,
緑矢印：「イスラム」$r = 0.54$,
青矢印：「バグダッド」$r = 0.49$,
黄矢印：「バント」$r = 0.48$,
桃矢印：「イスラム教」$r = 0.46$
緑ノード：カテゴリが国際関係の文書



(c) 第 3 ツリー（STAP 問題）
赤矢印：「都内」$r = 0.48$,
緑矢印：「笹井」$r = 0.45$,
青矢印：「幹細胞」$r = 0.44$,
黄矢印：「若山」$r = 0.44$,
桃矢印：「証拠」$r = 0.43$
赤，水ノード：カテゴリが社会，科学の文書

図 4: $\alpha = 0.23$ の TF に対するアノテーション結果