通時的な言語空間を用いた学術論文解析

Diachronical WordSpace Analysis of Academic Papers

中村 雄太 ‡　　浅野 泰仁 ‡
Yuta Nakamura　Yasuhito Asano
吉川 正俊 ‡
Masatoshi Yoshikawa

概 要

学術論文のトピックが時間的にどのように変遷したかを表す研究は，研究に欠かすことのできない先行研究の調査を手助けすることができると考えられる．先行研究では，bag of words でのモデル化や，LDA を用いた手法の提案が行われてきたが，これらの手法では，年代を経て主な意味が変化してしまった単語を正しく扱えていないという弱点がある．近年注目を集めている，word2vec に代表される単語埋め込みの手法の多くは，他の空間を比較することはできないが，変換を行うことで比較可能とし，時系列拡張することでそのような意味の変遷も考慮することができると考えられる．そこで，本研究は word2vec で獲得できる分散表現を用いたトピックの抽出とそれの変遷の可視化を試みる．

1. はじめに

学術論文は，単に知識を蓄積していくためのものではなく，過去の発見に基づいた研究をする際に必要な情報である．過去の学術論文を読み，先行研究を調査し当該分野の経年変化を含む全体像を把握することは，研究を進めていく中で欠かすことのできない行程である．しかし，先行研究の調査は，研究を始めて間もない人や自分の専門分野以外ではどこから手をつけていいのかわからず難航してしまうことがある．このような人に，当該の分野がどのようにこれまで変化してきたかを提示することは当該分野における全体像の把握に有用であると考えられる．

近年，DBLP*や arXiv†といった学術論文のデータベースの整備が進んでおり，このデータベースを活用する研究が盛んに行われている．このようなデータベースを活用する研究として，研究トピックの抽出 [1]，専門用語の検出 [2]，共著ネットワーク [3] の発見などが行われている．学術論文の理解を助ける研究の一つとして，研究トピックの抽出とその変遷を扱う研究 [1] があげられる．これは時間区分ごとにトピックを抽出し，その関係性を時間区間をまたいで同定することで，トピックがどのように変化していくかを検出する研究である．学術論文以外を対象にする場合でもこれらのトピックの抽出やその変遷を追う研究は数多くあり，Probabilistic Latent Semantic Indexing (p-LSI)[4] や Latent Dirichlet allocation (LDA)[5] を応用して行う Dynamic Topic Model[6] などがあげられる．

しかし，これらのどの研究においても，ある分野に関するトピックの変遷を知りたいとき，先行研究によって出力されたトピックとその変遷から，知りたい分野がどのトピックに属しているかユーザ自身が探し判断する必要があった．例えば，hadoop に関する分野について知りたい場合に，network や DB などのトピックの変遷が関係あるとユーザーが判断し，その内容を追わなければならなかった．これは，調べたい分野に詳しくない人は，どの単語が関連するかを判断することが難しいという問題点があった．また出てくる結果もそのトピックについての理解をすることができても，どのような概念と近くなったのかなど方向性については提示することができていないという欠点があり，ただの変遷だけではなく分野が，どのような方向性に向かっているかを含む全体像を俯瞰するという目的においては不十分であると言える．

また，これらの研究は単語を bag of words でモデル化しているため，似た単語を違うものとして扱うという問題点や，年代を経て単語自体の意味が変わってしまった場合でも同じ単語として扱ってしまうなど問題点がある．これらの問題は，word2vec[7] を各時間区分ごとに単語ベクトルを作成しそれらを変換行列を作成し結合する研究 [8, 9, 10] によって提案された手法を用いることで，解消できると期待される．クラスタリングを用いて word2vec の単語空間上にトピックを検出しそのトピックの方向ベクトルを定めることでトピック自体も同じ単語空間上に表すことができる．また，トピックをベクトルで表現することができるため，トピックがどのような単語と近い方向にどのくらいの速さで変遷しているのかを定量的に測るなど応用が期待される．

‡ 京都大学 情報学研究科,
Kyoto University Graduate School of Informatics

*http://dblp.uni-trier.de/
†https://arxiv.org/

本研究では，問合せ処理により知りたい分野の変遷を提示する手法を提案する．具体的な手法を以下に示す．まず，学術論文集合の各文章の題目とあらましを一つの文章とし，各時間区分ごとに word2vec によって学習しそれによって得られた単語ベクトルからその論文を表すベクトルを生成する．次に，文章集合を，問合せと近い概念を持った文章集合を選定し，次に出版の年度で文章集合を分割し，分割された集合内でクラスタリングを行いトピックを検出する．そして，異なった年度のトピック群同士を，分散表現で表されるトピックのベクトルを基に類似度を計算して結びつけ，そのトピック間の関係を同定し，トピックがどのように変化していっているのかを提示する．

本稿の構成を以下に示す．2. 節では，トピックの検出や word2vec に関する関連研究を述べる．3. 節では提案手法について述べる．4. 節では提案手法によって得られた実験結果をまとめ，それに対する考察を行う．5. 節で本論文のまとめと今後の課題，展望について述べる．

2. 関連研究

この章では，まず，トピックモデルやトピックの変遷の検出を試みている先行研究そして，word embedding のトピックの検出に関する先行研究について述べる．word2vec を時系列的に用いている研究や，類似する問題設定の先行研究について簡単にまとめる．

2.1 トピックの検出

トピックの検出は，これまで bag of words を用いて単語の局所表現であるところのベクトルを作りその後処理する手法が主であった．この手法は，単語の語順を考慮できていないことや似た単語の意味の扱いに問題があったが，簡単に文章をベクトル化できるために広く使用されていた．

近年 word2vec[7] の出現を受けて，分散表現を用いて得られる単語のベクトルを処理する手法が注目されるようになってきている．従来の手法と比べて，語順の問題や似た単語の扱いの問題を緩和できることが期待されている．さらに，局所表現であればベクトルの次元数は出現する単語数（10^4〜）となっていて，非常に大きな次元数の疎なベクトルとなっていたのに対して，分散表現で表されるベクトルは数百次元とより小さい次元で表すことができ，取り扱いやすいという特徴もある．

この節では，bag of words ベースのモデルのトピックとその変遷の検出と，word embedding のモデルを用いてトピックを扱う先行研究について記述する．

2.1.1 bag of words ベースモデル

bag of words を使用したトピックの検出モデルは，大きく三つに分けることができる．それぞれ，1. 確率モデル，2. グラフベースモデル，3. 行列分解モデルである．

確率モデルを利用したトピックモデルは，p-LSI[4] や LDA[5] に代表される．LDA は，トピックを文章に潜在的に分布しているものとして確率的にモデリングを行いその結果から，事前に与えられた数のトピックを抽出する手法である．LDA は様々な拡張がされている．LDA 自体の拡張としては，Teh ら [11] が事前にトピック数を指定する必要がなく，ドキュメント集合に応じて動的にトピック数を定める Hierarchical Dirichlet Process (HDP) を開発した．LDA や p-LSI を用いて学術論文集合のトピックを抽出する研究は多く行われており，著者の影響を考慮したトピックを算出するもの [12, 13] や，論文の引用情報を考慮に入れるもの [1] などがある．

そのほかに，グラフを用いた手法 [14] が存在する．この手法は，論文同士を使用している単語と引用関係からグラフを構成して，その単語がトピックとして使われる時その単語の引用グラフは密に接続されているという仮説を基に定式化をしている．これにより二つの単語の組からなるトピックを抽出することができている．

行列分解を用いた手法は近年より盛んに研究されるようになってきており，その中でも特に Non-negative Matrix Factorization (NMF) を用いたものが多くなってきている [9, 15]．NMF を用いた手法は，LDA などと比較して計算量が少なくて済むため，ソーシャルメディアやニュース記事などの即時性が求められるトピックの解析に用いられることが多い．

トピックそのものだけでなく，検出したトピックの変遷，つまり時間によってトピックの特徴語や構造がどのように変化していくかについての研究も盛んに行われている．トピックの変遷を検出する研究は判別手法と生成手法の二つに分けることができる．

判別手法は，文章集合におけるトピックを単語の組合せであると考えており，森永ら [16] は finite mixture model を使って，トピックの変遷がどのようにシフトしていくのかを，離散的に考慮している．

生成手法は，LDA の登場により活発になっている手法であり，Blei らは [6] は LDA を拡張し，トピックの盛衰をもモデル化してパラメータを学習する Dynamic Topic Model を提案した．Wang ら [17] は LDA を拡張した Topic Over Time (TOT) という手法を開発した．この手法は，トピックの変遷を年度などの離散的な時間区分ではなく連続的に扱う手法であり，論文のデータやメールのデータを用いて検証を行われている．これまで

は，SNS などのデータであっても一日ごとなどに分割して使用しなければならず，その分割が正しいのかという疑問があり，この手法は連続時間で扱えるようにしたことでこの問題を緩和した．

その他にも，He ら [1] は，学術論文により適したモデルを開発した．これまでの研究では学術論文をデータとしている研究であっても，引用情報をその論文との関係性も考慮した上で用いているものはなく，論文では引用情報は大切な要素の一つであるという考えから，引用情報を加味したモデルを作成した．また，引用情報はすべて同じ重要度ではないと考えて，その割合についても学習するモデルを LDA を拡張して作成してる．

2.1.2 word embedding モデル

word embedding を使用したモデルは，word2vec や doc2vec の拡張であり，bag of words ベースのモデルとは違い低次元であり計算量が比較的少なく済み，トピック自体もベクトル化して，単語ベクトルと同列に扱うことができるという特徴がある．

Niu ら [18] は事前に LDA で文章集合を学習させておき，それによってつけられたラベルを word2vec の学習時に付与することで，単語とトピックのベクトルを学習する Topic2vec を開発した．Li ら [19] は，トピックと単語を同時に学習するモデルを作成した．これらの手法は bag of words を用いてきたこれまでの手法と同程度またはそれ以上の成果を出しており新しい方向性として期待されている．Batmanghelich ら [20] は LDA のトピック数を動的に決める Hierarchical Dirichlet Process (HDP)[21] を拡張し，トピックを学習しそのそのトピックを単一のベクトルとして扱うのではなく，単語空間を norm が 1 になるように正規化することで超球面上のある方向ベクトルとして扱い，von Mises-Fisher 分布を用いて表現することで，言葉の意味も考慮したトピックモデリングが可能になるとしている．

2.2 word embedding の時系列適応

異なるパラメータで通常学習した単語空間は相互に比較することができないが，それを可能にし，時系列的に単語空間を扱うという手法が近年研究されている．Vaca ら [9] はまずコーパスを時間単位ごとに区切り，最初の時間単位は通常通り Skipgram Negative Sampling （SGNS）で学習した後，次の時間単位の学習の初期パラメータを前の時間単位の単語空間とすることで連続的に扱い，複数年たっても比較することができるとし，それによって言葉が経年でどのように変化していったのかを検出する手法を提案した．Hamilton ら [8] はこの研究に加えて，複数の年度のベクトルを直交行列を用いて変換することで複数の単語空間の比較を可能にし，経年でどのような単語が意味の変遷が起こりやすいかなど意味の変化と単語の性質について検証した．Yao ら [22] は上記二つの研究を踏まえて検証し，時間区分を超えて代替物となるもの，例えば 2016 年の Obama が 1993 の Clinton と近いということを発見できるとしている．また，Zhang ら [10] らは，変換行列を用いて複数年度の単語空間を比較し，1980 年代の Walkman と 2010 年代の iPod など似た側面のある二つのものが，Sony や Apple などどのような違いがあるかを発見する手法を提案した．

3. 提案手法

この章では，提案手法を説明する．まず，問題設定を行い，その後に論文集合に対しての分散表現の獲得手法について記述する．次に，ユーザから与えられた問合せの処理について扱い，そして，それを用いたトピックの抽出方法とその変遷の検出方法について扱う．

3.1 問題設定

本研究は分散表現を用いるため，単語集合の分散表現は N 次元のベクトルで表される．本研究では全ての分散表現は長さ 1 に正規化されているものとする．単語分散表現系列を $\mathbf{W} = \langle w_1, \cdots, w_V \rangle$ と表す．本研究では，二つの単語の分散表現 w_i と w_j の間の類似度を式 (1) のコサイン類似度を用いて算出することとする．

$$\mathrm{sim}(w_i, w_j) = \frac{w_i \cdot w_j}{||w_i||\,||w_j||} \tag{1}$$

$D = \{d_1, \cdots, d_l\}$ を学術論文の集合とする．本研究ではこの学術論文の集合も単語と同様に，N 次元の分散表現空間で表す．学術論文集合の分散表現は $\mathbf{D} = \{\mathbf{d_1}, \cdots, \mathbf{d_l}\}$ で表すことができる．文章ベクトル $\mathbf{d_i}$ は，全ての単語集合 V と，TD-IDF によって得られる重み TFIDF_{d_i} と，文章にその単語が存在するかを表す関数 ζ(3) を用いて，長さ 1 に正規化する関数 normalize を用いて式 (2) のように表される．

$$\mathbf{d_i} = \mathrm{normalize}\left(\sum_v^V \zeta_v^i \cdot \mathrm{TFIDF}_{w_v} \cdot w_v\right) \tag{2}$$

$$\zeta_v^i = \begin{cases} 1 & v \in d_i \\ 0 & otherwise \end{cases} \tag{3}$$

ただし，関数 normalize は式 (4) と表される．

$$\mathrm{normalize}(v) = \frac{v}{||v||} \tag{4}$$

この時各文章間および各文章および単語間の類似度は，コサイン類似度を用いて算出することとする．

本研究ではトピックの検出のために，von Mises-Fisher (vMF) 分布を用いる．vMF 分布は半径 1 の超球面上の

ガウス分布であり，主に方向データのモデリングに使用される．角度によって類似性を測ることができるため，単語分空間のモデル化を行うことができると考えられる．また，vMF は中心ベクトル μ とその広がり κ という二つのパラメータを持つ．トピックは単一のベクトルとしてではなくある程度広がりを持つと考えられるため，トピックを vMF を用いて分類することは適切であると考えられる．そこで，本研究では vMF 混合分布によって文章ベクトルをクラスタリングし，そのクラスタをトピックとすることとする．ある文章 d_t が，トピックをあらわす vMF から生成される確率密度分布 $p(d_t|\Theta)$ をパラメータ群 Θ を用いて式 (5),(6),(7) で表す．

$$p(d_t|\Theta) = \sum_{m=1}^{M} \pi_m f(d_t|\mu_m, \kappa_m) \tag{5}$$

$$f(d_t|\mu_m, \kappa_m) = C_N(\kappa) \exp(\kappa_m \mu_m^T w_t) \tag{6}$$

$$C_N(\kappa) = \frac{\kappa^{N/2-1}}{(2\pi)^{N/2} I_{N/2-1}(\kappa)} \tag{7}$$

このとき，M は混合数とし，π を異なる vMF の重みとし，I は第一種変形 Bessel 関数を表す．本研究では簡単のため，vMF を用いたクラスタリングによって得られた二つのトピック間の類似度を中心ベクトルのコサイン類似度を用いて測ることとする．

文章集合 D を，出版された時刻 $t(\in [1, n])$ を用いて互いに素な集合 $D(1), \cdots, D(n)$ に分割する．ただし，$D = \cup_{t=1}^{n} D(t)$ を満たす．

3.2 分散表現の獲得手法

word embedding を利用した分散表現の獲得は，近年注目を集めている．本研究では word2vec を用いて，単語や文章の分散表現を獲得する．word2vec によって得られた分散表現は，足し算や引き算といった演算をすることができ，国と首都の関係などベクトルでその関係性を表すことができるとされている．

複数の時間区分の分散表現を獲得するとき，全ての時間区分から一つの分散表現を獲得する手法と，複数の分散表現を獲得しそれらを変換する二つの手法が存在する．前者は，学習したものを変換無しに正しく比較することができるという特徴があるが，例えば “cloud” という言葉の使われ方が 1990 年と 2015 年で違うように，言葉の主な使われ方が変遷してしまった場合に正しくとらえられないという弱点がある．その一方複数の分散表現を獲得しそれらを変換する手法はこれらの問題を緩和することはできるが，異なるパラメータで学習した単語空間同士は通常比べることができないため別の単語空間に変換を行わなければならないという弱点がある．本研究では，Vaca ら [9] や，Hamilton ら [8] の研究を基に，まずはじめの時間区分 $D(1)$ のコーパスで学習しその結果を初期状態として次のコーパスを学習させ，分散表現を連続的に学習させた後，式 (8) のような変換行列 R を作成した．

$$\mathbf{R}^{(t)} = \underset{\mathbf{Q}^T\mathbf{Q}=I}{\arg\min} \left\| \mathbf{W}^{(t)}\mathbf{Q} - \mathbf{W}^{(t+1)} \right\|_F \tag{8}$$

ここで，$\mathbf{W}^{(t)}$ は，時間区分 t で学習した分散表現であり，$\mathbf{Q}$ は直交行列である．また，$\mathbf{W}^{(t)}$ には全ての単語を使わずに，出現回数の多い単語はその使われ方が変わらないと仮定し，最も出現回数の多い単語 1000 語を用いた．本研究では，出版年度を基に時間区分を定義することし，その粒度は一年とする．

word2vec は，単語の分散表現を獲得する手法であり，文章の分散表現は単語の分散表現から獲得するものとする．本研究では，文章集合 D 内の文章 d_i の分散表現を獲得し，それを用いてトピックの分類を行う．我々は，論文においてその論文を最もよく表現している部分が題目とあらましであると仮定し，d_i の分散表現は題目とあらましの分散表現で近似されると考えた．本研究では，文章ベクトルは，題目とあらましを一つの文章として，各時間区分で TF-IDF 値を計算しその重みを利用して、単語ベクトルの重み付き平均を文章ベクトルとしている．

また，単語の処理に関しては，単純に文章を空白文字で分割したものよりも，n-gram をとって，フレーズとして使用した方が理解がしやすい場合がある．しかし，すべての n-gram をとるのは語彙空間が大きくなりすぎてしまうという欠点がある．そこで，本研究では，mikolov ら [7] に則り，頻度に応じた n-gram を作成することにする．この手法の二つの連続している単語 w_i, w_j が 2-gram として連結させるべきかについての評価関数を式 (9) で表す．

$$\mathrm{score}(w_i, w_j) = \frac{\mathrm{count}(w_i \cdot w_j) - \delta}{\mathrm{count}(w_i) \times \mathrm{count}(w_j)} \tag{9}$$

δ は多くの単語連結ができてしまうことを防ぐためのパラメータであり，この評価値が閾値を超えた場合に，単語を連結する．また，この手法を複数回繰り返すことで，2-gram から 3-gram を作り出すこともできる．本研究では，“natural language processing” など三語程度までまとまりのある言葉があるととらえるのが妥当であると考えたため上記の手法で 3-gram まで作成している．なおそのほかの単語の処理として，名詞・動詞・形容詞・副詞のみを全て原型に戻して小文字にして取り出し，英単語 (a-z) 以外のものはすべて排除するという処理を行っている．

3.3 問合せ処理

この節では，3.2 節で獲得した分散表現 $\mathbf{D}$ から問合せを考慮したベクトル集合を抽出する手法について説明する．本研究での問合せ q とは，$q \in \mathbf{W}$ を満たす単問合せのことを指す．問合せとして与えられた言葉に関するトピックがどのように変遷していくかを抽出することが目的であるため，単語 q だけでなく，時間区分 t も入力とする．つまり，q_t という表記で，時間区分 t の単語 q を意味するものとする．これにより，単語の分散表現 $w_{q_t} \in \mathbf{W}^{(t)}$ を定めることができ，これに近いものを検索結果とする．このようにすることで，例えば 2015 年の “topic modeling” と直接対応するものが 2000 年になくても “text classification” を検出できるなど年代を超えて近い概念を検索することができる．検索結果は通常ある閾値を超えたものや，最も類似度が高い k 件を正解集合とするものが多いが，新しい概念などは昔は存在しておらず今は非常に数が多くなることが予想され，閾値ベースだと新しい時間区分での問合せ結果が多すぎたり，上位 k 件だと古い時間区分で遠い文章が問合せ結果に含まれてしまう可能性があるため，本研究ではこれらを組み合わせてある閾値以上のかつ上位 k 件とする．

3.4 トピックの抽出

この節では，3.2 節および 3.3 節で獲得した分散表現 D からトピックを抽出する手法について説明する．文章の分散表現である D は N 次元空間上に分布しており互いに比較が可能であり，同じトピックは近くの距離に存在しクラスタを形成しており，距離が遠いクラスタは別のトピックであると考えることができる．そのため，我々はトピックの抽出手法としてクラスタリングを用いた．また，単語の分散表現は長さを 1 に正規化し，超球面上に分布させることで方向データとして扱うため，ここで混合 vMF でモデリングする．vMF のパラメータは EM アルゴリズムで最尤推定する．EM アルゴリズムの E ステップにおける負荷率の計算には，ハード方式とソフト方式が存在 [23] し，ソフト方式が厳密解であるが本研究では速度の都合上ハード方式を使用している．

また，通常混合分布の混合数ははじめに与える必要があるが，ある文章集合の正しい混合数つまり，トピック数を人手で与えることは難しいため，Bayes Information Criterion （BIC）を用いて混合数を決定することがあり [24]，本研究においても BIC で混合数を決定する．BIC は尤度関数 L，混合数 k，データのサンプル数 n を用いて式 (10) のように表される．

$$\mathrm{BIC} = -2 \cdot \ln(L) + k \ln(n) \tag{10}$$

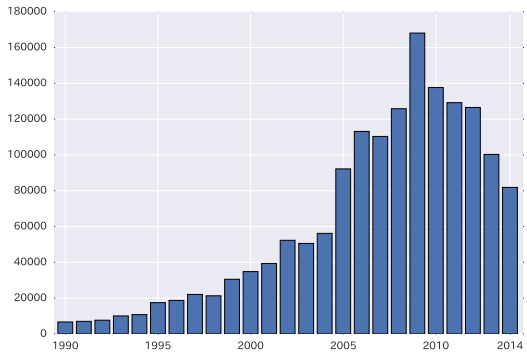

図 1: 各年の論文数

3.5 トピックの変遷の検出

トピックの変遷とは，複数の時間区間にわたり，出現しているトピックを特定しその内容の変遷を検出するものである．3.4 節で求めたトピックは区間 t の学術論文集合である $D(t)$ に対して求めたものであり，区間 t 以外の情報は使用していない．トピックの変遷の検出は，異なる二つの時間区間 t, t' のトピックとの関係性を解析して行う．本研究では簡単のため隣り合った年代 t と $t+1$ のトピックの間で類似度を計算し，最も類似度が高いものを同一のトピックと見なす．

4. 評価実験

本研究では，提案手法の有効性を確かめるために，Aminer[25] によって提供されている Citation Network のデータセットを利用した．本研究では各論文のタイトルとあらましの双方が必要となるため，このデータセットのうちタイトルとあらましのどちらにもかけがない文章のうち，データの数が比較的多くなる 1990 年から 2014 年までを対象とした．本研究の対象となっている論文数は合計で 1,570,902 件の論文を対象としている．各年の論文の数を図 1 に示す．

4.1 実験結果

実験環境は，Ubuntu 15.04，Intel core i7 6770k，Memory 64GB の環境で行った．実験に用いたコードは Python2.7.12 で実行され，word2vec および n-gram を用いたフレーズ作成，そのほか言語処理は gensim 0.13.2 を用いた．

この節では，提案手法による問合せを与えた場合のトピックの変遷の検出についての有効性について評価するために，まず word2vec を用いて時間区分を考慮した分散表現の獲得したことにより単語の変遷を追えていることを確認した後，問合せ処理に使われている年代を超えた代替物を検索することができるという性質を確認する．

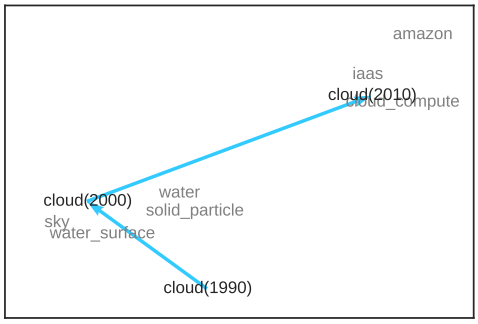

図 2: cloud という単語の変遷の検出

表 1: hadoop の代替物検索

年度	似ている単語
1990	mimd, sun, sprite, vax, share_memory_multiprocessor
2000	network_workstation, pvm, linux_cluster, disk_array, file_server

表 2: 2012,2013 年のトピック

トピック番号	似ている単語
1	topic_modeling, latent_dirichlet_allocation_lda
2	text, document

これらを踏まえて，トピックが年代を経てどのように変化していったかを検証する．

4.1.1 単語の変遷の検知

本研究で学習したモデルが，先行研究と同様に単語の意味の変遷を検出できていることを検証する．ここではケーススタディーとして，“cloud” を用いる．この言葉は，クラウドコンピューティングという言葉が出てくる前は，主な意味は空に浮かぶ雲として利用されていた．つまり，空の雲から，クラウドコンピューティングという意味へ変移していることが図示できれば良い．word2vec などの単語空間を低次元に圧縮して可視化するときに t-Stochastic Neighbor Embedding (t-SNE)[26] がよく用いられる．t-SNE は，Principal Component Analysis (PCA) と同様に次元削減の手法を用いて，確率分布を用いて二次元または三次元に可視化するため手法であり，高次元のデータを構造を失わずに可視化することができることで知られている．ここでは，“cloud” という単語が 2014 年の単語の中をどのように移動しているかを図示するために，tSNE を用いて次元圧縮しこれらを図示したものが図 2 である．通常 tSNE はある程度多くの単語数がある方がうまく可視化することができるため，図 2 以外に，最も出現頻度の高い 500 単語も入れて tSNE で圧縮し，“cloud” に関係あるところのみ図示している．この図によると，元々は sky などと近かったが IaaS などと近い概念になっていることがわかり，適切に図示されていることがわかる．

4.1.2 年代を超え類似する単語の検知

次に，問合せ処理にも使用されている，時代を超え類似する単語の検出について検証する．ここではケーススタディーとして “hadoop” という単語を用いる．2014 年の “hadoop” という単語に割り当てられたベクトルと最も近い単語を 1990 年，2000 年で探したものが表 1 となる．hadoop とは大規模データを分散処理するフレームワークであり代替物には分散処理や分散ファイルシステムなど直接の代替物がない年代においても近い単語が検索できていることがわかる．この性質により，どの単語が関連するのかを時代を超えて検出することができると考えられる．

4.1.3 トピックの変遷

トピックベクトルの変遷を検証する．ここではケーススタディーとして，“topic_modeling” という単語を用いる．2014 年の “topic_modeling” という単語と近い単語を 3.3 節で定義した手法でクエリ処理を行う．ここでは，コサイン類似度の閾値を 0.4 とする．問合せ処理によって得られた論文のベクトル集合を，平滑化するため連続二年をコーパスとして扱い，vMF によってクラスタリングを行いトピックを抽出する．2012 年，2013 年の論文集合をコーパスとして抽出されたトピックベクトルと似ている単語を表 2 に示す．なお，BIC に最適なトピック数を算出したところ 2 つになっている．

このうちトピック番号 2 に着目する．このトピックはソースとなるドキュメントなどを対象にしているトピックであると考えられる．これがどのように変遷しているかを示したのが図 3 となる．問合せ処理により，出力された件数が 10 件を超える 1997 年 1998 年以降に絞って図示を行っている．wikipedia などから，マイクロブログなど shorttext に変遷していることが図示されている．

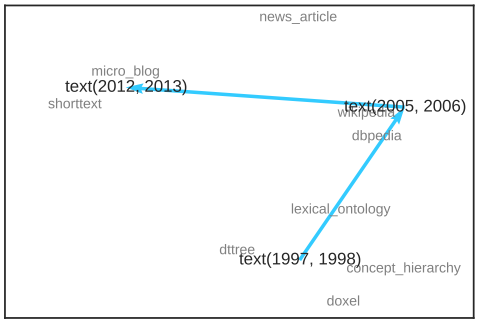

図 3: トピックの変遷

5. むすびに

本研究では，分散表現を用いて，学術論文のトピックの変遷を提示する手法を提案した．本手法は，問合せ処理も可能であるため欲しい概念に関する変遷を獲得することができ，時系列的な分散表現を使用しているため，単語の使われ方が変化したり新しい概念が現れた場合にも対応できていると考えられる．

今後の課題としては，まず本研究では評価として少ない数のケーススタディしか行えていないため他のケースでの検証や，既存の研究との比較を行っていくことが挙げられる．現在は過去の変化の方向を可視化することにとどまっているが，ベクトルでトピックを表すことができているので，変遷の速度による分野の成熟度や，変遷に一定の法則などが見いだせるのであれば近い未来どのように変遷していく可能性があるのかを提示することができないか検討していきたい．さらに，現在は一つのトピックのみを扱っているが，複数のトピックの関連性も定性的に測ることができるため，それらを用いて複数のトピックの関連性についても提示することができるように検討していきたい．

参考文献

[1] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. Detecting topic evolution in scientific literature: how can citations help? In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 957–966. ACM, 2009.

[2] Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. Terminology extraction: an analysis of linguistic and statistical approaches. In *Knowledge mining*, pp. 255–279. Springer, 2005.

[3] Xiaoming Liu, Johan Bollen, Michael L Nelson, and Herbert Van de Sompel. Co-authorship networks in the digital library research community. *Information processing & management*, Vol. 41, No. 6, pp. 1462–1480, 2005.

[4] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 289–296. Morgan Kaufmann Publishers Inc., 1999.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.

[6] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120. ACM, 2006.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[8] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*, 2016.

[9] Carmen K Vaca, Amin Mantrach, Alejandro Jaimes, and Marco Saerens. A time-based collective factorization for topic discovery and monitoring in news. In *Proceedings of the 23rd international conference on World wide web*, pp. 527–538. ACM, 2014.

[10] Yating Zhang, Adam Jatowt, and Katsumi Tanaka. Towards understanding word embeddings: Automatically explaining similarity of terms. In *Big Data (Big Data), 2016 IEEE International Conference on*, pp. 823–832. IEEE, 2016.

[11] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 2012.

[12] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 306–315. ACM, 2004.

[13] Ding Zhou, Xiang Ji, Hongyuan Zha, and C Lee Giles. Topic evolution and social interactions: how authors effect research. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 248–257. ACM, 2006.

[14] Yookyung Jo, Carl Lagoze, and C Lee Giles. Detecting research topics via the correlation between graphs and texts. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 370–379. ACM, 2007.

[15] Janani Kalyanam, Amin Mantrach, Diego Saez-Trumper, Hossein Vahabi, and Gert Lanckriet. Leveraging social context for modeling topic evolution. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 517–526. ACM, 2015.

[16] Satoshi Morinaga and Kenji Yamanishi. Tracking dynamics of topic trends using a finite mixture model. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 811–816. ACM, 2004.

[17] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 424–433. ACM, 2006.

[18] Liqiang Niu, Xinyu Dai, Jianbing Zhang, and Jiajun Chen. Topic2vec: Learning distributed representations of topics. In *Proceedings of the 2015 International Conference on Asian Language Processing (IALP)*, pp. 193–196. IEEE, 2015.

[19] Shaohua Li, Tat-Seng Chua, Jun Zhu, and Chunyan Miao. Generative topic embedding: a continuous representation of documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 666–675, 2016.

[20] Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Sam Gershman. Nonparametric spherical topic modeling with word embeddings. *arXiv preprint arXiv:1604.00126*, 2016.

[21] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pp. 1385–1392, 2005.

[22] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Discovery of evolving semantics through dynamic word embedding learning. *arXiv preprint arXiv:1703.00607*, 2017.

[23] Kurt Hornik and Bettina Grün. movmf: An r package for fitting mixtures of von mises-fisher distributions. *Journal of Statistical Software*, Vol. 58, No. 10, pp. 1–31, 2014.

[24] Scott Chen and Ponani Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. darpa broadcast news transcription and understanding workshop*, Vol. 8, pp. 127–132. Virginia, USA, 1998.

[25] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, pp. 990–998, 2008.

[26] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, Vol. 9, pp. 2579–2605, Nov 2008.