分散表現空間解析モデルに基づく研究トレンドに関する考察

中村　雄太† 　浅野　泰仁† 　吉川　正俊†

† 京都大学 情報学研究科　〒 606-8501 京都府京都市左京区吉田本町

E-mail: †y-nakamura@db.soc.i.kyoto-u.ac.jp, ††{asano,yoshikawa}@i.kyoto-u.ac.jp

あらまし　学術論文のトピックやその変遷を検出する研究は近年盛んに研究されており，研究トピックの変遷は研究者が研究テーマを決める際の重要な情報である．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，これらの手法は語の順番や似た単語の扱いに問題があった．近年注目を集めている，word2vec に代表される単語埋め込みの応用である doc2vec では，文章の分散表現を獲得することができ，語順や似た単語の問題を緩和でき，トピック単位だけではなく論文単位で変遷を追うことができると期待される．そこで，本研究は doc2vec で獲得できる分散表現を用いたトピックの抽出とその変遷の検出方法について検討する．

キーワード　doc2vec, word embeding, 研究トピック

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研究トレンドを把握することは，研究テーマを決め，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．先行研究の調査にはトピックやキーワードを用いられることが多く，初心者や専門外の分野を調べるときは，この適切なトピックやキーワードがわからず難航してしまうと考えられる．適切なトピックやキーワードを用いて調査することができれば，調査の速度があがり，適切な理解が得ることができるであろう．

近年，DBLP(注1) や arXiv(注2) 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．研究トピックの抽出の応用研究の一つに，研究トピックの変遷の検出がある．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分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のかを検出する研究である．これらの，研究トピックやその変遷がわかれば，当該分野の研究トレンドやトピックおよびキーワードを俯瞰することができ，初心者や専門外の分野の調査に役立つことが期待される．このため，近年トピックの抽出やその変遷の研究は活発に行われている．

研究トピックの変遷に関する研究は，情報検索の分野で多くの先行研究があり，その多くが bag of words を用いてモデル化を行っている．これらの研究は，Probabilistic Latent Semantic Indexing (p-LSI) [4] や Latent Dirichlet allocation (LDA) [5] などの確率モデルに基づくものが最も一般的であり，そのほかに文章-単語分布行列に対して行列分解を行うものや，グラフ構造を用いたものがある．

しかし，これらのモデルは bag of words を用いてモデル化を行っているため，語の順番が考慮されていないという問題や，似た意味を持つ単語が別の単語として扱われてしまうという問題があった．近年これらの問題を緩和することが期待される，word2vec [6] などに見られる word embedding を用いた分散表現の学習に注目が集まっている．その中でも，word2vec の応用である，doc2vec [7] は文章の分散表現を獲得できる手法として知られている．

また，大きなトピックやその変動だけでなく，特定の単語に関するトピックおよびその変遷を知ることができればよりユーザの理解を支援することができると考えられる．しかし，これまでの研究 [1] では，既にできているトピックとその変遷からその単語があるトピックがどこにあるかをユーザ自身が探し判断する必要があった．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．具体的には，分散処理に関するトピックの大きな変動を知ることができ，その特徴語として現れた hadoop という言葉の周りのトピックがどのように生まれ変遷していくかについて知りたい場合に，ユーザ自身がその大きなトピックの中をみて判断する必要があった．これは，ユーザ問合せを受け付け，問合せと近い文章集合からトピックおよびその変遷を検出することができると考えられる．

そこで，本研究では，問合せ処理をすることができる doc2vec を用いた研究トピックとその変遷である，研究トレンドを検出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，doc2vec を用いてその学術論文一つ一つの分散表現を獲得する．次に，文章集合を，問合せ処理を行うことで，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，分散表現を基に類似度を計

(注1)：http://dblp.uni-trier.de/

(注2)：https://arxiv.org/

算して結びつけ，そのトピック間の関係を同定し，トピックの変遷を検出し，最後にそれぞれのトピックに適切なラベルを与える．

本稿の構成を以下に示す．2. 節では，トピックの検出やその変遷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 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．トピックの変遷の検出を試みている先行研究や，似たような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，簡単にベクトル化できるために広く使用されていた．

近年 word2vec [6] 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，スパースで非常に大きな次元数の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 ベースのモデルと，word embedding のモデルによるトピックの抽出についての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物である．この 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 [8] 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．

この 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 [9], [10] 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 [11] 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 [12], [13]. 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．

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同列に扱うことができるという特徴がある．

Niu ら [14] 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 [15] 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．

2.2 トピックの変遷の検出

近年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が盛んである．トピックの変遷を検出する研究は 2.1.1 節で述べた bag of words ベースモデルで主に研究されており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，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，トピックの変遷の時間を連続的に使用するという手法を用いて，論文のデータやメールのデータを用いて検証を行った．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．

これらのどの手法も，ある言葉に関するトピックの検出およびその変遷の検出，つまり問合せに対応することができていなかった．例えば，LDA を用いた手法であれば，結果の特徴語に

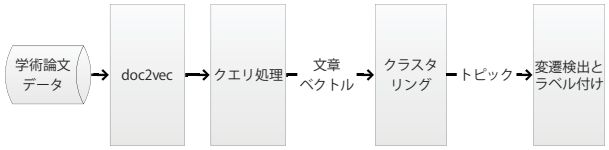

図 1　提案手法の全体像

問合せまたは問合せと近い意味の単語が出現しているかをユーザが調べ，そのトピックの変遷を追う必要があった．しかし，その分野に詳しくない場合は問合せとどの単語が近い意味かどうかを判別することができないという問題があった．そこで本研究では，ユーザが自身で判断するのではなく，問合せを与えることで分散表現を用いて問合せと近い文章を見つけ，それらを文章集合としてトピックおよびその変遷を検出する手法を提案する．

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．まず提案手法の全体像を図 1 に示す．

3.1 問 題 設 定

$\mathbb{W} = \{w_1, w_2, \cdots, w_V\}$ を単語集合とする．bag of words を用いたモデルの場合は，これを V 次元のベクトルとして処理するが，本研究は分散表現を用いるため，N 次元のベクトルへ関数 $\mathbf{f} : \mathbb{W} \to \mathbb{R}^N$ を用いて写像して使用している．単語集合の分散表現は，$\mathbf{W} = \langle \mathbf{f}(w_1), \cdots, \mathbf{f}(w_V) \rangle$ というように表すことができる．本研究では，二つの単語の分散表現 $\mathbf{f}(w_i)$ と $\mathbf{f}(w_j)$ の間の類似度をコサイン類似度を式 (1) で算出することとする．

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

$D = \{d_1, \cdots, d_m\}$ 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に写像する．写像関数 $\mathbf{f_d} : D \to \mathbb{R}^N$ を用いて，学術論文集合の分散表現は $\mathbf{D} = \langle \mathbf{f_d}(d_1), \cdots, \mathbf{f_d}(d_m) \rangle$ で表すことができる．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．

トピック集合 $\mathbf{z} = \langle z_1, \cdots, z_k \rangle$ は，文章集合 D の部分集合で表され，トピックの数を k とし，それぞれのトピックの集合の要素数を l_z として，二つのトピック $\mathbf{z}_i$ と $\mathbf{z}_j$ の類似度は式 (2) で表される群間平均で求められると定義する．

$$\mathrm{clustersim}(\mathbf{z}_i, \mathbf{z}_j) = \frac{1}{l_{zi} \cdot l_{zj}} \sum_{k_1}^{l_{zi}} \sum_{k_2}^{l_{zj}} \mathrm{sim}(\mathbf{z}_{ik_1}, \mathbf{z}_{jk_2}) \qquad (2)$$

二つのクラスタ内部の類似度の算出方法は，この他にも最短距離や最長距離など様々な手法があるが，これらの手法はこの類似度を利用してクラスタリングする際に，空間濃縮などの問題を起こす可能性があることで知られている．そこで，このような問題が起こらず簡潔に表現できる群間平均を用いた．

トピックの変遷を検出するため，文章集合 D を，その時間 $t(\in [1, n])$ 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．区間 t での文章集合 $D(t)$ のトピック集合を $\mathbf{z}(t)$ とすると，トピックの変遷とは，区間 t と t' のトピック集合 $\mathbf{z}(t)$ と $\mathbf{z}(t')$ のそれぞれのトピック間の関係性を，“同じトピック”，“類似しているトピック”，“関係なし” という三つから同定することである．

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，bag of words を用いたモデルが抱えていた問題を緩和することができると期待され，近年急速に注目を集めている．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，$d_i(t)$ 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．

次に，分散表現の獲得手法であるが，本研究では doc2vec を用いる．doc2vec は，単語の分散表現と同時に文章の分散表現を獲得することができる．我々は，まず論文の題目とあらましを一つの文章としてならべ，これをデータセットとして doc2vec で学習を行っている．

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，まとまり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 [6] 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 (3) で表す．

$$\mathrm{score}(w_i, w_j) = \frac{\mathrm{count}(w_i \cdot w_j) - \delta}{\mathrm{count}(w_i) \times \mathrm{count}(w_j)} \qquad (3)$$

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 Processin” 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，すべて小文字にし，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 $\mathbf{f}$ 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，$q \in \mathbb{W}$ を満たす単問合せのことを指す．本研究での問合せ処理は，問合せとして与えられた言葉の周辺のトピックとその変遷を抽出することが目的であるため，単純に単語を含むか含まないかではなく，その単語を含んでいる文章およびその文章と類似度が最も高い K 件の論文の集合とする．問合せ q を含んでいる文章集合を $\mathbf{D}_q = \langle d_{q1}, d_{q2}, \cdots, d_{qn} \rangle$ とする．問合せ q を含む文章 $d_{qj}(j = 1, 2, \cdots, n)$ と最も高い類似

度を持つ K 件の論文集号を，$\mathrm{topKsim}(\mathrm{d_{qj}})$ と表記する．ただし，$\mathrm{topKsim}(d_{qj}) \subset \mathbf{D}$ である．ここで，本研究で問合せ処理によって得られる文章集合 $\mathbf{D}_q$ は式 (4) で表される．

$$\mathbf{D}_q = \bigcup_{i=1}^{n} \mathrm{topKsim}(d_{qi}) \tag{4}$$

またこの文章集合に対応する分散表現集合を $\mathbf{f}_q$ と表すこととする．

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 $\mathbf{f}_q$ 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 $\mathbf{f}_q$ 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．クラスタリングには，大きく階層的クラスタリングと非階層的クラスタリングの二つに分けることができる．非階層的クラスタリングは，k-means など事前にトピック数を指定しなければならない一方で，階層的クラスタリングは，事前にトピック数を指定することなく，群と群の類似度がある閾値を超えるかでクラスタリングを行う手法である．学術論文集号の正しいトピック数を知ることはできないため，事前にトピック数を与える必要がない階層的クラスタリングを用いることとした．階層的クラスタリングの類似度の計算手法には，3.1 節のクラスタ間の距離と同様に，複数の手法が提案されているが，空間濃縮などが起こらずコサイン類似度を計算することができる群間平均法を用いることとする．

3.5 ラベルの付与

トピックやその変遷を提示する際には，そのトピックに所属する文章を羅列するのではなく，特徴語に基づくラベルがあるとユーザの理解を支援することができる．LDA や p-LSI を扱う論文集合があった場合，これをユーザに提示するときに “topic model” というラベルを付与することが例としてあげられる．そこで，本研究では，3.4 節で抽出したトピックにラベルを付与する．適したラベルとは，他のクラスタに含まれないが，自分のクラスタに多く含まれているそのクラスタの特徴語であると考えることができる．そこで，トピックに所属する文章を一つの文章として結合し，トピックの数の個数だけ文章を用意し doc2vec と同様に事前に n-gram をとり前処理を文章集合に，TF-IDF 法を用いてその単語の重要度を算出した．また，TF-IDF に代表される bag of words の手法では文章に重要でない単語があると精度が落ちるため，文章の内容に大きく寄与する，名詞，形容詞，副詞，動詞のみを用い，また動詞は原型に戻して使用している．最終的に用いるラベルは，これらの TF-IDF 法によって各単語に与えられた重要度が最も高い 5 単語とする．

3.6 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，この $\mathbf{z}(t)$ と $\mathbf{z}(t')(t' < t)$ との関係性を解析して行う．

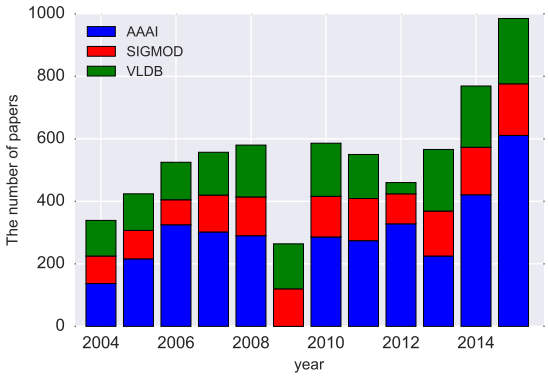

図 2　各年の論文数

区間 t のトピック $\mathbf{z}_i(t)$ と区間 $t(< t')$ のトピック $\mathbf{z}_j(t')$ との関係性について，He ら [1] と同様に，二つのパラメータ ϵ_1, ϵ_2 を用いて，以下の三つの関係性を定める．

同じトピック: $\mathrm{clustersim}(\mathbf{z}_i(t), \mathbf{z}_j(t')) \geqq \epsilon_1$

似たトピック: $\epsilon_1 > \mathrm{clustersim}(\mathbf{z}_i(t), \mathbf{z}_j(t')) \geqq \epsilon_2$

関係のないトピック: $\mathrm{clustersim}(\mathbf{z}_i(t), \mathbf{z}_j(t')) < \epsilon_2$

この二つのパラメータは，実験的に求めるものとする．

この分析の結果，同じトピックであると考えられていたものの内容がどのように変遷しているのか，或いは途切れたのかを検出し，トピックの変遷の検出を試みる．

4. 実　　験

4.1 データセット

本研究では，提案手法の有効性を確かめるために，三つの学会の学術論文データを用いた．三つの学会とは，データベース系の VLDB，SIGMOD と自然言語・人工知能学会の AAAI である．それぞれの学会のデータを各 2004 年から 2015 年（AAAI は 2009 年は非開催）のものを取得している．このうち，収集した PDF から各論文のタイトルとあらましを抽出し，そのどちらにもエラーが起こらなかった 6,605 論文を対象としている．各年の論文の数を図 2 に示す．

4.2 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，doc2vec 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，クラスタリングおよび可視化には SciPy 0.17.0 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 doc2vec を用いて分散表現の獲得を行った結果を確認し，次にそれを用いて問合せ処理を行わずにトピック検出の結果について扱い，そしてそれらを用いたトピックとその変遷の検出について説明する．最後に，問合せを与えた場合のトピックの変遷の検出について説明する．

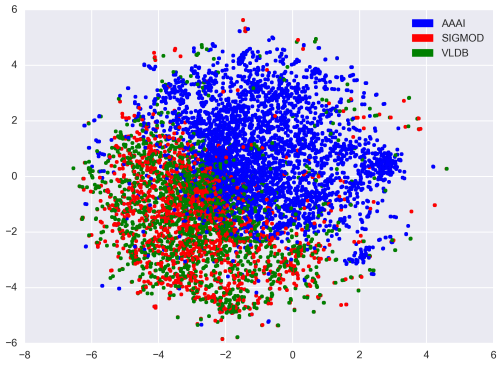

図 3　分散表現空間の可視化

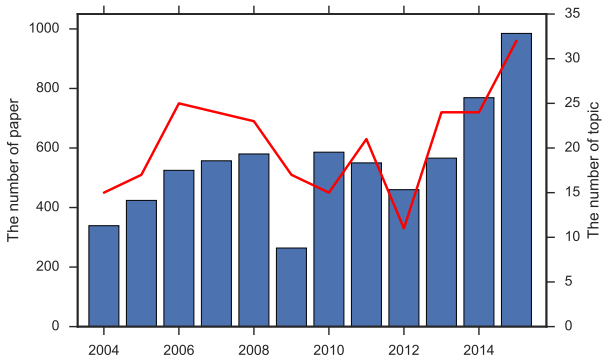

図 4　各年度の論文数とトピック数

4.2.1　doc2vec を用いた分散表現の獲得

本研究で使用している文章数は，合計で 6,605 件とそれほど多くないため，通常 doc2vec を使う際は 10〜20 程度である繰り返しの回数を 30 にして実験した．それ以外のパラメータは gensim の初期値から変更は行っていない．すべてのデータを doc2vec にかけ，得られた分散表現空間を t-Stochastic Neighbor Embedding (t-SNE) [18] を用いて図 3 のように可視化した．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．

図 3 を見ると，データベース系の学会である，SIGMOD と VLDB は点が重なり合っているが，この二つの学会と人工知能や自然言語処理を扱う AAAI はあまり重なりがなく，学会ごとの特徴が論文内の言葉から取得できていることがわかる．

4.2.2　トピックの検出

4.2.1 節で得られた分散表現からトピックの抽出階層的クラスタリングを用いて行う．階層的クラスタリングには，クラスタとして認識する閾値をパラメータとして持っている．今回は各年度のクラスタ数を 20 前後で表すことが適切であると考えられたため，各年度でクラスタリングを行う際に，その年度の階層間の距離のうち最大のものに 0.92 をかけたものを閾値とした．これによって得られたクラスタの数を図 4 に示す．おおむね論文数に応じてトピックの数が増減していることがわかる．

表 1　提案手法と LDA の比較

手法	トピックの特徴語	論文数
提案手法	1: word, topic_model, video, entity, tweet	168
	2: workload, ml, join, architecture, memory	144
	3: lowrank, learn, sample, label, spurious_sample	102
	4: path, decomposition, pi, diff, constraint	80
	5: agent, voter, coalition, election, mechanism	64
LDA	1: game, solve, text, word, topic	102
	2: join, database, query, processing, performance	77
	3: performance, application, query, machine, assignment	64
	4: query, domain, item, feature, graph	58
	5: belief, robot, game, character, security	49

次に，実際に検出されたトピックを用いたケーススタディを行う．2015 年のデータを用いて得られたクラスタの中でクラスタの規模が大きい 5 個のトピックとその特徴語，そしてそのクラスタの大きさを表 1 に示す．比較として，LDA を用いて得られた結果も記載する．LDA は，2015 年のデータのタイトルとあらましから，名詞，動詞，形容詞，副詞だけを抜き出しそれを原型に変換して bag of words を用いてベクトル化したものを用いて学習を行っている．ハイパーパラメータは，提案手法と比較するために，トピック数を提案手法によって得られた 27 個に固定し，繰返し回数を 20 回に設定した．そのほかのハイパーパラメータは gensim の初期値のままである．トピックの特徴語は，尤度が高い単語の上位 5 件を特徴語であると考えた．また，LDA を用いたトピックの論文数の検出方法は，与えた論文に対して最も尤度が高いトピックに帰属するという手法をとっている．

この結果，所属する論文の数は提案手法と LDA で違いが見られるが，トピックモデルに対応するトピックが提案手法でも LDA でも最も大きトピックとして考えられるなど近いトピックを抽出することができていると考えることができる．一方で，LDA では，performance や game など同じ単語が複数のトピックに上がってきており，また抽出されている単語も，assignment などすぐにどの分野かわかるようなものではない．その一方で，提案手法では，n-gram をとっているため，topic_model など人間にとってわかりやすいラベルが付与できていることが確認できる．

4.2.3　トピックの変遷の検出

トピックの変遷の検出は，隣り合う二つの年度のトピックの類似度を計算することで，その関係性を同定する．得られる三つの区分の割合が同じようになるように，“同じトピック” の閾値 ϵ_1 を 0.1 とし，“類似しているトピック” の閾値 ϵ_2 を 0.08 とした．この結果各年度の同じトピック数，類似しているトピック数の分布は図 5 のようになった．

次に，提案手法によるトピックの変遷の検出についてのケー

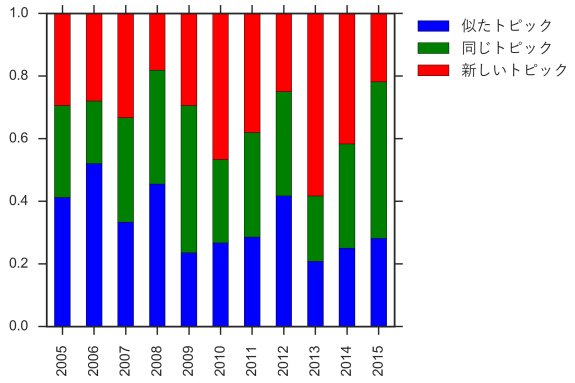

図 5　トピック間の関係の割合

表 2　問合せ hadoop を与えた時の 2004 年のトピック

特徴語	論文数
page, storage, write, logstructure, layout	2
dataow, parallel, operator, flux, failure	2
data_warehouse, operational, warehouse, model, chal	2

ススタディーとして，hadoop に代表される分散処理についてのクラスタを例示する．このクラスタのトピックの変遷は図 6 のようになる．得られたトピックの変遷によると，hadoop 関連のクラスタは 2005 年頃から発生しており，もともとは mysql などのデータベースから始まり，data_management に変わり，2010 年に hadoop という言葉が出てきている．また，2011 年，2012 年頃は mapreduce という言葉が上がってきていたが，2013 年から transaction や tenant などに移行し，2015 年には ML が特徴語として上がっていることが見て取れる．これは，実際に古くはデータベースに始まりそこから mapreduce が生まれ，それが変化していき分散トランザクションを処理するようになり機械学習に応用されるようになった，hadoop などに代表される分散処理の歴史を表せていると考えられる．

4.2.4　問合せを与えたときの変遷の検出

提案手法による問合せを与えたときのトピックの変遷の検出についてのケーススタディーとして，hadoop を問合せとして与えたときの結果を記す．4.2.3 節のケーススタディを見ることで得られた結果によって hadoop に関連する分野を俯瞰することはできたが，細かい変遷までは追うことができなかった．そこで，提案手法で hadoop という問合せを与えた時の例を確認して，細かい変遷を追うことができているかどうかを評価する．

まず hadoop という問合せを与えたときに，hadoop という言葉は含まないが，その周辺分野の論文としては適切なものを取得できているかについて評価する．そのために，hadoop という言葉が現れる前である 2004 年のトピックを図 2 に示す．なお，問合せを与えた場合は，より詳細にトピックが分かれていることが望ましいため，クラスタリングの数が 15 個程度になるようにクラスタリングの際に用いる閾値を 0.75 に調整している．hadoop という単語は 2008 年から現れた物であり，単純に問合せを含む論文のみを取得する手法であれば 2008 年以前の文章を取得することはできないが，提案手法では分散表現を用いて似た文章を取得しているので，概念が現れる以前の文章も取得することができていることが確認できる．

表 3　問合せ hadoop を与えたときの 2015 年のトピック

特徴語	論文数
realtime, reef, heron, cloud, stack	9
hive, gobblin, spark, tez, relational	6
interactive, tableau, statistic, workflow, way	5
dream, purity, byteslice, scan, rdf	5

次に，hadoop という問合せを与えた時の 2015 年のクラスタの大きさが 5 以上のトピックを表 3 に示す．問合せを考慮しない場合であれば，特徴語 worload, ml, join, architecture, memory で構成されていたが，問合せを考慮したことにより，上から順に realtime streaming を扱うトピックと hive などデータベースを扱うトピック，Business Intelligence を扱うトピック，RDF エンジンを扱うトピックがその中に存在していたことがわかる．heron や byteslice など一つの論文内で提唱されているシステムが特徴語として現れてしまっているが，トピックの大きさが少ないため起きていることであると考えられる．

最後に，これらのトピックのつながりの一部を図 7 に例示する．この一例は，分析を考慮した分散処理の変遷を表していると考えられ，レンジ問合せの高速化からストリーム処理へと変遷していることが見て取れる．問合せを与えていない場合と同様にトピックおよびその変遷が取得できており，小さなトピックであっても変遷を検知できていると考えることができる．

5. むすび

本研究では，分散表現を用いたトピックとその変遷の検出方法を提案した．また，それだけでなくこれまでの研究ではできなかった問合せ処理に対応させた．実験によりトピックおよびその変遷の検出，問合せ処理によるより詳細なトピックおよびその変遷を検出できることが確認された．

今後の課題としては，まず本研究では評価としてケーススタディしか行えていないため他のデータセットに適用することや，既存の研究との比較を行っていくことが上げられる．また，現在の手法は文章や単語が分散表現を用いて同じ空間にあるという特性を生かし切れていない部分がある．今後トピックの変遷を，類似度が非常に高い論文がどのように空間上を動いていったかをベクトルベースで表し，その次にどのような変化を起こすかの予測や，変化の内容をその変化したベクトルから推定できるようにすることなどが考えられる．さらに，現在は論文というデータセットを扱う上で考慮することを欠かすことができない引用情報を考慮できていないという欠点がある．今後はこの引用情報も含めて考慮していきたい．また，今後の展望として，これらの結果を利用して研究トレンドを先読みし，研究テーマの推薦などより役に立つシステムの構築を心がけていきたい．

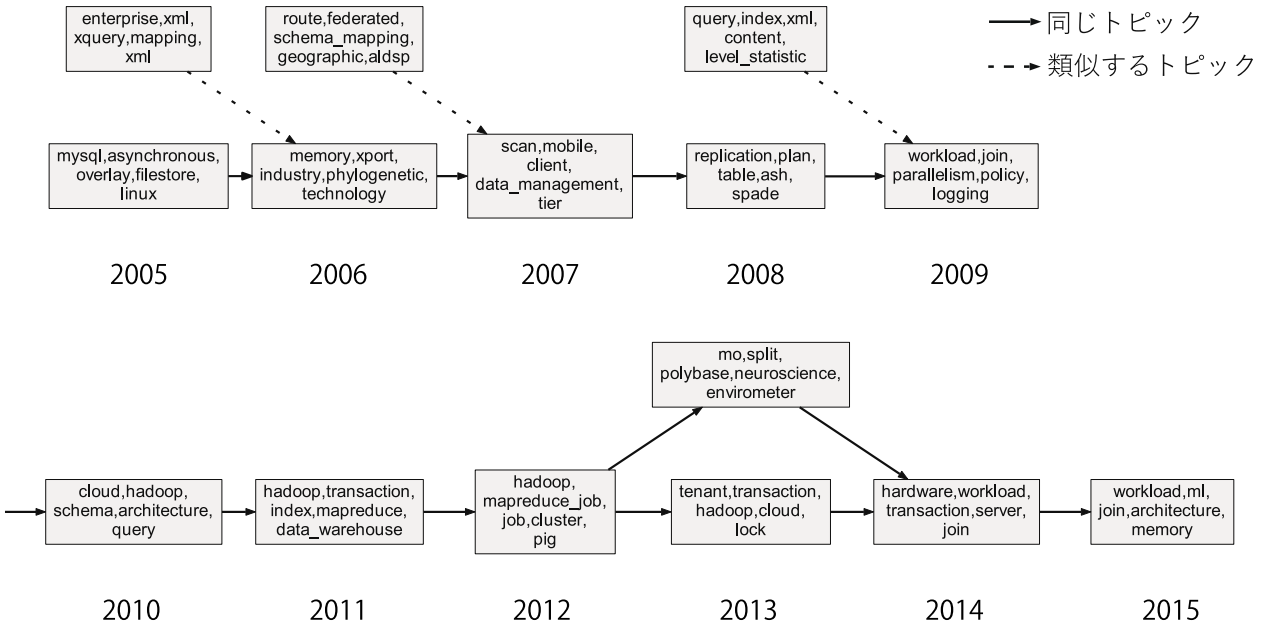

図 6　分散処理のトピックの変遷

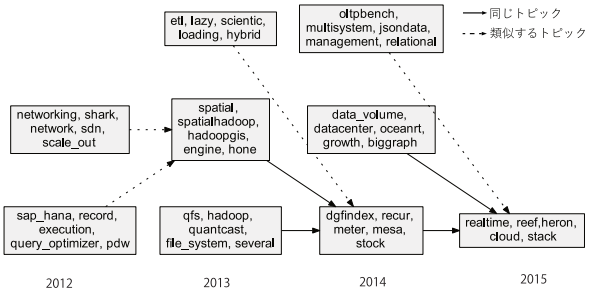

図 7　問合せ hadoop を与えた時のトピックの変遷の一部

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] 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.

[7] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, Vol. 14, pp. 1188–1196, 2014.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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