DBSJ Japanese Journal Vol. 14-J, No. 13
March 2016

引用情報を利用した論文データベースにおけるトピックの変遷の検出

Detecting Topic Evolution in Bibliographic Databases Exploiting Citations

伊藤 寛祥♡ 天笠 俊之♢ 北川 博之♠

Hiroyoshi ITO Toshiyuki AMAGASA
Hiroyuki Kitagawa

近年，論文データベースには膨大な論文が集約されており，そこから情報を抽出する手法は注目を集めている．中でも論文データベースにおけるトピックの抽出，およびトピックの変遷の検出に関する手法は，時間の経過によってある研究トピックがどのように発生，分岐，統合，消滅したかを知るための良い手がかりとなるため，研究者にとって有用な情報となる．本研究では，行列分解に基づくトピックの抽出および時間経過によるトピック変遷の検出手法を提案する．特に論文の本文に加えて，引用情報を考慮したトピックの抽出，および変遷を検出する新しい手法を提案する．さらに評価実験によりその有効性を検証する．

In recent years, a large amount of data have been accumulated in bibliographic datases, and hence mining techniques for such data have been attrancting great attentions. Among them, it has become more important to extract topics and their chronological evolution process. If reserchers get evolutions of research topics, they can obtain useful information, e.g. , how research topics have emerged, diversified, integrated, or diminished as time passes. In this research, we propose a scheme for detecting the chronological evolution process in a bibliography database based on matrix factorization. More precisely, we exploit citations to improve the accuracy due to the fact that they convey useful information. We conduct experiments to show the feasibility of the proposed scheme.

1. はじめに

近年，学術分野において論文データを中心としたリポジトリの整備が進んでいる．コンピュータサイエンスでは DBLP[1]，宇宙物理学分野では ADS[2] や arXiv [1]，医学分野では PUPMED[3]，などが著名である．このような論文データベースに対して分析を行うことにより，各分野における研究トピックの抽出や，主要論文の抽出，研究者コミュニティの抽出などが可能となる [12, 9, 13, 6]．なかでも，研究トピックの変遷は各分野の発展過程をデータベースから抽出する技術であり，近年よく研究されている [9, 3]．

一方，文書データにおけるトピック検出は，活発に研究がなされており，p-LSI [4]，LDA [2] など確率モデルに基づく研究が多くなされている．これに対して，近年では，文書-単語分布行列に対して Non-negative Matrix Factorization（NMF）を適用することにより，文書中のトピックを検出する手法が注目されている [15]．

そこで，本研究では NMF に基づき論文データベースからトピックの変遷を抽出する手法を提案する．具体的には，文書集合を重複した時間区間ごとに分割し，引用情報を考慮した上で特徴抽出を行う．得られた特徴に対して，NMF を適用し，トピック検出を行う．さらに，隣接する時間区間におけるトピック群同士を，共通する文書を手掛かりに関連付け，トピックの変遷を検出する．

本論文の構成は以下のとおりである．第 2 節で提案手法のベースとなる NMF の概要，数学的性質について触れる．第 3 節で本研究の提案手法について述べる．第 4 節で実験結果について述べ，提案手法で検出されたトピックの変遷について議論を行う．第 5 節で本研究と関連する研究について述べ，第 6 節で本論文のまとめと今後の課題，展望を述べる．

2. Non-negative Matrix Factorization

Non-negative Matrix Factorization（NMF）とは非負値行列分解のことで，入力として与えられた非負行列を二つの非負行列に分解する手法である．この手法を入力データ行列に対して適用すると，データセット内の頻出パターンを抽出することができる．NMF は主に画像認識における特徴抽出や，音声信号処理におけるスペクトル分析などに用いられている手法である [17, 11]．

文書集合に対して NMF を適用する場合，各文書をベクトルに変換し，それを並べたものを入力データ行列として扱う．したがって入力行列 X の大きさは，$N_d \times N_f$ となる．ここで N_d は文書集合における文書数，N_f は各文書の特徴量の数となり，ここでは文書を bag-of-words にする際の単語数である．そして入力データ行列 X を，$X \approx WH$ を満たす，非負行列 W，H に近似的に分解する（図 1）．ここで，出力される行列 W は $N_d \times K$，H は $K \times N_f$ となる．ここで K はトピックの数であり，一般に N_f より非常に小さい値をとるように設定する．

しかしながら，NMF は解析的に行うことができないため，損失関数を設定し，それを最適化することで分解を行うのが一般的である．式 (1) は損失関数の一例である．以下の式を $W \geq 0$，$H \geq 0$ という条件のもとで最適化することで行列分解できる．

$$L = \arg\min_{W,H} \|X - WH\|_F^2 \tag{1}$$

$\|\cdot\|_F^2$：フロベニウスノルム

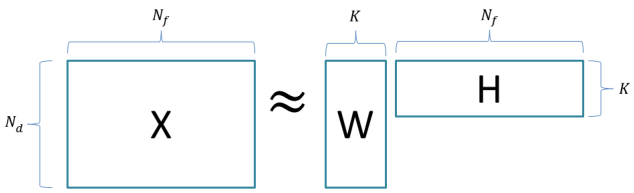

図 1: NMF のイメージ

出力として得られる行列 W，H のもつ性質について述べる．H は各トピックの単語分布を表現する行列となる．$H = (\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_k)^T$ とした場合，各ベクトル $\mathbf{h}_i$ は次元数 N_f のベクトルであり，これがすなわちトピックとなる．したがって入力文書集合のトピックを抽出するという文脈においては，行列 H を推定することが目的

♡ 学生会員　筑波大学システム情報工学研究科
hiro.3188@kde.cs.tsukuba.ac.jp
♢ 正会員　筑波大学システム情報系
amagasa@cs.tsukuba.ac.jp
♠ 正会員　筑波大学システム情報系
kitagawa@cs.tsukuba.ac.jp

[1] http://dblp.uni-trier.de/
[2] http://adswww.harvard.edu/
[3] http://www.ncbi.nlm.nih.gov/pubmed

となる．このとき抽出される各トピック $\mathbf{h_i}$ は，特徴（単語）空間において，文書が集中して存在している方向ベクトルとなる．したがって，NMF におけるトピックはクラスタリングに関しても有効であると言うことができる [16]．W は各文書が持つトピックの割合を表現する行列となる．$W = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_{N_d})^T$ とした場合，各ベクトル $\mathbf{w_j}$ は次元数 K のベクトルとなり，このベクトルの各要素の大きさが，すなわち文書 j の持つ各トピックの大きさとなる．ここで，要素の大きさに閾値を設定することで，文書集合を K 個のクラスタにソフトクラスタリングすることが可能である [16]．また論文集合におけるトピック k の強さを式 (2) で算出することができる．

$$TopicIntensity_k = \sum_{i}^{N_d} W_{ik} / \sum_{i}^{N_d} \sum_{j}^{K} W_{ij} \tag{2}$$

W_{ij} : 行列 W の i, j 成分

3. 提案手法

本論文では NMF をベースに用いた，トピックの変遷を検出する手法を提案する．本手法の対象とする論文データは論文情報として，タイトル，アブストラクト，タイムスタンプ，引用情報を保持しているものとし，すべてのタイムスタンプにおける論文データセットを D とする．手法のおおまかな流れは，まずトピックの変遷を検出する際の時間的粒度を決定し，その時間区分で論文の集合 D を分割，分割された文書集合をそれぞれ行列に変換する．その後，その行列それぞれに行列分解を適用することで，各時間区分におけるトピックを抽出する．最後に前後の時間区分において抽出されたトピック間を結びつけることでトピックの変遷を検出する．以下の節でそれぞれの処理について具体的に記述していく．

3.1 論文データセットの分割と行列への変換

提案手法では，時間経過によるトピックの変遷を検出するため，データセット D をある一定の期間で分割する．その際，前後の時間区分におけるトピックが滑らかに接続されるように時間区分同士をオーバーラップさせる（図 2）．こうしてできた時間区分 t における論文集合を $D_X^{(t)}$ とする．

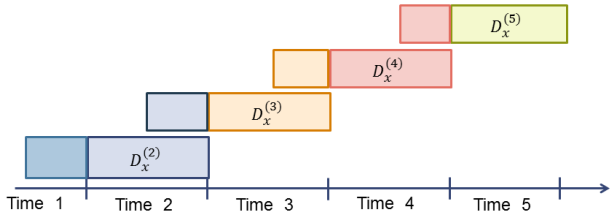

図 2: 時間区分をオーバーラップさせるイメージ

時間区分で分割された文書集合を行列に変換する．各文書の特徴量は bag-of-words とし，TF-IDF 等の手法でベクトル化する．また，文書のコンテンツはタイトル，アブストラクトとする．こうして時間区分 t における論文集合 $D_X^{(t)}$ を行列に変換したものを $X^{(t)}$ とする．このとき行列 $X^{(t)}$ の大きさは $N_d^{(t)} \times N_f$ となる（$N_d^{(t)}$ は時間区分 t における文書数，N_f は単語数）．

ここで，論文におけるトピック抽出において，引用情報は重要な手がかりになる点に着目する．そこで，ある論文から引用されている論文の情報も文書行列に組み込むことを考える．すなわち，行列 $X^{(t)}$ における論文が引用している論文の集合 D_C^t を $X^{(t)}$ と同様に行列に変換し，得られた行列を $C^{(t)}$ とする．このとき行列 $C^{(t)}$ のサイズは $N_c^{(t)} \times N_f$ となる（$N_c^{(t)}$ は時間区分 t における論文が引用している論文数）．

3.2 トピックの抽出

この節では，論文の集合からトピックを抽出する行列分解を提案する．ある論文で引用されている論文は，引用元の論文と共通のトピックを持っている可能性が高いと考えられる．そこで，本論文では引用情報を考慮した行列分解を行う．具体的には，論文行列 $X^{(t)}$ と被引用論文行列 $C^{(t)}$ を結合した行列に対して NMF を適用する（図 3）．行列分解は式 (3) に示す損失関数の最適化により行う．

$$L = \underset{W_X^{(t)}, W_C^{(t)}, H^{(t)}}{\arg\min} \|X^{(t)} - W_X^{(t)} H^{(t)}\|_F^2 + \delta \|C^{(t)} - W_C^{(t)} H^{(t)}\|_F^2 \tag{3}$$

パラメータ δ によって被引用論文がトピックに与える影響力を調節できる．パラメータ δ が 0 に近づくほど被引用論文がトピックに与える影響力は小さくなり，∞ に近づくほど被引用論文がトピックに与える影響力は大きくなる．

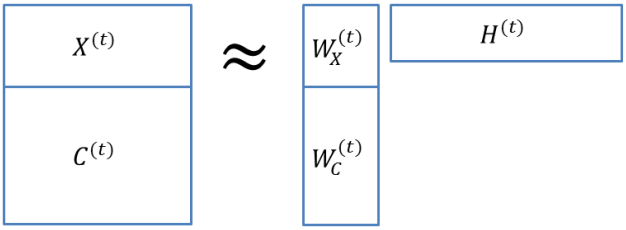

図 3: 被引用論文を考慮した行列分解

出力される各行列の大きさは $W_X^{(t)}$ は $N_d^{(t)} \times K$，$W_C^{(t)}$ は $N_c^{(t)} \times K$，$H^{(t)}$ は $K \times N_f$ である．ここで K は抽出されるトピックの数であり，トピック数が小さいほどより大域的，トピック数が大きいほどより局所的なトピックが抽出される．

式 (3) は $W_X^{(t)}, W_C^{(t)}, H^{(t)}$ に関して同時に凸にはならない．しかしながら，[7] で提案される更新式を適用することで，局所最適解を求めることができる．損失関数を最適化する際の KKT 条件は以下である．

$$W_X^{(t)} \geq \mathbf{0}, \quad W_C^{(t)} \geq \mathbf{0}, \quad H^{(t)} \geq \mathbf{0} \tag{4}$$

$$\nabla_{W_X^{(t)}} L \geq \mathbf{0}, \quad \nabla_{W_C^{(t)}} L \geq \mathbf{0}, \quad \nabla_{H^{(t)}} L \geq \mathbf{0} \tag{5}$$

$$W_X^{(t)} \odot \nabla_{W_X^{(t)}} L = \mathbf{0}, \quad W_C^{(t)} \odot \nabla_{W_C^{(t)}} L = \mathbf{0}, \quad H^{(t)} \odot \nabla_{H^{(t)}} L = \mathbf{0} \tag{6}$$

ここで $\odot$ はアダマール積である．式 (3) の各行列に関する勾配は以下である．

$$\nabla_{W_X^{(t)}} L = W_X^{(t)} H^{(t)} H^{(t)^T} - X^{(t)} H^{(t)^T} \tag{7}$$

$$\nabla_{W_C^{(t)}} L = W_C^{(t)} H^{(t)} H^{(t)^T} - C^{(t)} H^{(t)^T} \tag{8}$$

$$\nabla_{H^{(t)}} L = (W_X^{(t)^T} W_X^{(t)} H^{(t)} + \delta W_C^{(t)^T} W_C^{(t)} H^{(t)}) - (W_X^{(t)^T} X^{(t)} + \delta W_C^{(t)^T} C^{(t)}) \tag{9}$$

式 (6) に勾配を代入することにより更新式を導出した．

$$W_X^{(t)} \leftarrow W_X^{(t)} \odot \frac{[X^{(t)} H^{(t)^T}]}{[W_X^{(t)} H^{(t)} H^{(t)^T}]} \tag{10}$$

$$W_C^{(t)} \leftarrow W_C^{(t)} \odot \frac{[C^{(t)} H^{(t)^T}]}{[W_C^{(t)} H^{(t)} H^{(t)^T}]} \tag{11}$$

$$H^{(t)} \leftarrow H^{(t)} \odot \frac{[W_X^{(t)^T} X^{(t)} + \delta W_C^{(t)^T} C^{(t)}]}{[W_X^{(t)^T} W_X^{(t)} H^{(t)} + \delta W_C^{(t)^T} W_C^{(t)} H^{(t)}]} \tag{12}$$

式 (10), (11), (12) を繰り返し適用することにより損失関数は減少していき，ある程度収束したところで行列分解が完了する．以上のアルゴリズムを Algorithm1 に示す．

出力された行列 $W_X^{(t)}$ は各文書が持つトピックの割合を表現する行列である．ここで，その割合に閾値を設けることでトピックベースのクラスタリングを行う．時間区分 t のトピック i に分類される，オーバーラップした期間における論文集合を $T_i^{(t)}$ であらわす．以上の集合は式 (13) で定式化される．

$$T_i^{(t)} = \left\{ d \in D_X^{(t-1)} \cap D_X^{(t)} \middle| \frac{W_{X_{di}}^{(t)}}{\sum_j W_{X_{dj}}^{(t)}} \geq \theta \right\} \tag{13}$$

$W_{X_{di}}^{(t)}$: 行列 $W_X^{(t)}$における文書 d のトピック i の成分

Algorithm 1 引用情報を考慮した行列分解

Input: $X^{(t)}, C^{(t)}, \delta, \epsilon$

Output: $H^{(t)}, W_X^{(t)}, W_C^{(t)}$

1: $H^{(t)}, W_X^{(t)}, W_C(t) \leftarrow$ random non-negative init
2: $\epsilon' \leftarrow maxInt$
3: $\epsilon \leftarrow \frac{\epsilon'}{2}$
4: $\delta \leftarrow$to choose in [0,∞]
5: **while** $abs(\epsilon' - \epsilon) \geq \epsilon$ **do**
6: $\quad W_X^{(t)} \leftarrow W_X^{(t)} \odot \frac{[X^{(t)} H^{(t)^T}]}{[W_X^{(t)} H^{(t)} H^{(t)^T}]}$
7: $\quad W_C^{(t)} \leftarrow W_C^{(t)} \odot \frac{[C^{(t)} H^{(t)^T}]}{[W_C^{(t)} H^{(t)} H^{(t)^T}]}$
8: $\quad H^{(t)} \leftarrow H^{(t)} \odot \frac{[W_X^{(t)^T} X^{(t)} + \delta W_C^{(t)^T} C^{(t)}]}{[W_X^{(t)^T} W_X^{(t)} H^{(t)} + \delta W_C^{(t)^T} W_C^{(t)} H^{(t)}]}$
9: $\quad \epsilon' \leftarrow \epsilon$
10: $\quad \epsilon \leftarrow L(X^{(t)}, W_X^{(t)}, W_C^{(t)}, H^{(t)})$
11: **end while**

3.3 トピック間の類似度の算出

3.2 節では各時間区分におけるトピックおよび各トピックに関連する文書集合が抽出された．次に，隣接する時間区分におけるトピックについて，それが十分類似していれば，同じトピックであると判定する．このとき，トピック間の類似度を $sim\left(\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t-1)}\right)$ とする．ここで $\mathbf{h}_i^{(t)}$ は時間区分 t における i 番目のトピックである．

ここで，トピックの変遷に意味付けを行うために類似度の分類を行う．トピック間の類似度は (1) 同じトピック，(2) 似ているトピック，(3) 違うトピック，という 3 パターンで意味付けできると考えられる．そこで，求められるトピック間の類似度に閾値を設けることで，以上の分類を行う．それぞれの類似度は以下で定式化する．

- $\mathbf{h}_i^{(t)}$ と $\mathbf{h}_j^{(t-1)}$ は同じトピック

$$sim\left(\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t-1)}\right) > \epsilon_1 \tag{14}$$

- $\mathbf{h}_i^{(t)}$ と $\mathbf{h}_j^{(t-1)}$ は似ているトピック

$$\epsilon_1 \geq sim\left(\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t-1)}\right) > \epsilon_2 \tag{15}$$

- $\mathbf{h}_i^{(t)}$ と $\mathbf{h}_j^{(t-1)}$ は違うトピック

$$\epsilon_2 \geq sim\left(\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t-1)}\right) \tag{16}$$

ここで各パラメータは $\epsilon_1 \geq \epsilon_2 \geq 1/K$ を満たすように設定する．

類似トピックの判定には，トピックベクトルの類似性やトピックに含まれる共通文書を利用することが考えられる．以下でそれぞれについて定式化する．

トピックベクトルの類似度はコサイン類似度で測ることができる [8]．したがってトピックベクトルによる類似度を以下で定式化する．

$$\begin{aligned} sim\left(\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t-1)}\right) &= \cos\left(\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t-1)}\right) \\ &= \frac{\mathbf{h}_i^{(t)} \cdot \mathbf{h}_j^{(t-1)}}{|\mathbf{h}_i^{(t)}||\mathbf{h}_j^{(t-1)}|} \end{aligned} \tag{17}$$

しかしながら，トピックベクトルの類似度は，単語分布の類似性であるため，人間の感性に近い接続が行われない場合が想定される．本研究ではトピックに含まれる共通文書を利用した接続方法を提案する．具体的にはトピックに分類される文書集合の Jaccard 係数の大きさを接続の強さとし，以下のように定式化する．

$$\begin{aligned} sim\left(\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t-1)}\right) &= Jaccard(\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t-1)}) \\ &= \frac{|T_i^{(t)} \cap T_j^{(t-1)}|}{|T_i^{(t)} \cup T_j^{(t-1)}|} \end{aligned} \tag{18}$$

本研究では以上の 2 つの類似度算出法による接続の差異について実験で議論する．

4. 実験

4.1 データセット

本研究では，手法の評価のために，論文データベース arXiv [1] を対象に手法を適用した．arXiv は物理学を中心とし，他に数学や計算機科学の論文を保存，公開している web サイトである．論文データ取得のためにクローラを制作し，1995 年から 2014 年までのすべての論文，およびそれらの論文が引用している論文を取得した．図 4 は各年の論文数のグラフ，図 5 は各年の被引用論文数のグラフである．時間経過とともに論文数はともに増加傾向にある．また，図 6 は，ある年の論文数に対する被引用論文数の遷移を表したグラフで，論文 1 件に対する被引用論文数は平均約 7.38 倍である．

本実験は 64bit Ubuntu14.04，Intel Core i7@3.70GHz コア数 8，32GB RAM で構成された PC 上で行った．提案手法の実装は Python2.7.6 で行った．データセットの行列への変換はライブラリ gensim0.10.2 を使用し，行列計算にはライブラリ numpy1.8.2, scipy0.13.3 を使用した．

実験において設定した各パラメータについて述べる．トピックを抽出する際のトピック数は $K = 50$，論文のクラスタリングにおける閾値は $\theta = 0.16$，類似度の閾値は，Jaccard 係数で接続する場合は $\epsilon_1 = 0.5$，$\epsilon_2 = 0.1$，コサイン類似度で接続する場合は $\epsilon_1 = 0.5$，$\epsilon_2 = 0.25$ である．トピックを抽出する期間は年単位とし，時間区分をオーバーラップさせる期間は 4ヶ月とする．

4.2 トピックの変遷の検出

図 7 は本手法で検出されたトピックの変遷の一部である．図中における四角形は本手法で検出されたトピックを表し，内部にトピックを形成する単語が記載されている．また各単語の大きさは本手法によって検出された各トピックにおける単語の強さであり，人の手によるラベリング等は行っていない．2012 年におけるトピック “hole, black, horizon, accret, entropi, gravit, charg, solut” を起点として，隣接する時間区分のトピックの中から類似度の大きいものを探索して結合させたものである．線の濃さは 3.3 節で

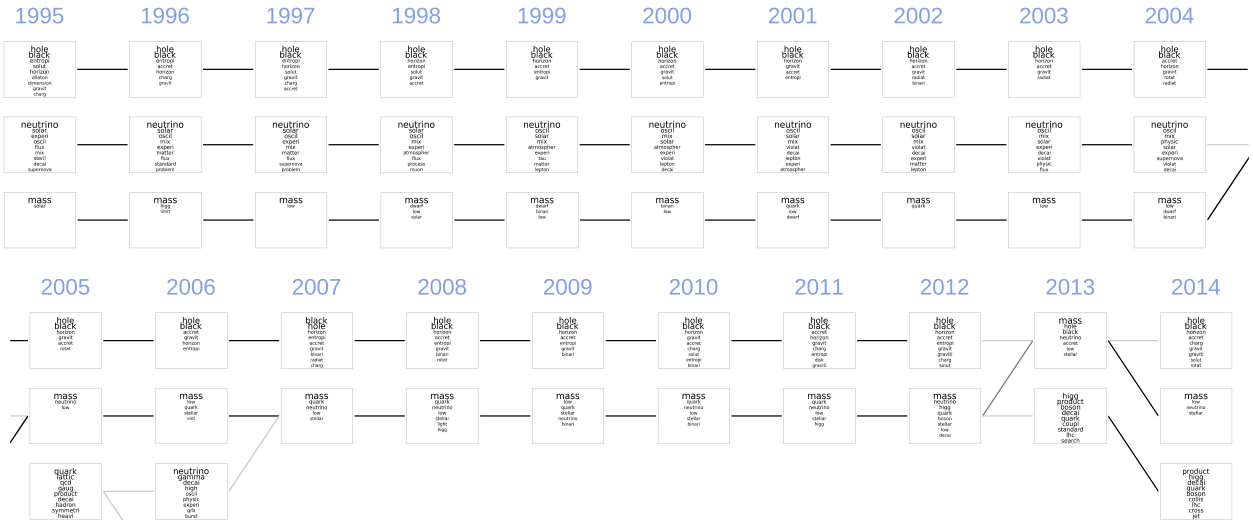

図 7: 本手法で検出されたトピックの変遷の一部

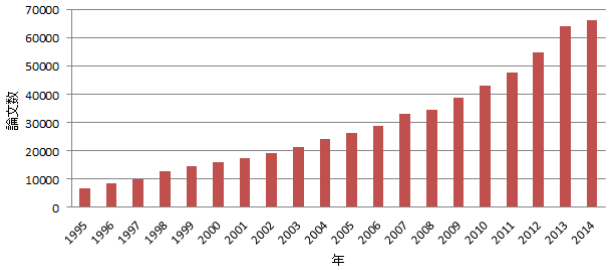

図 4: 各年の論文数

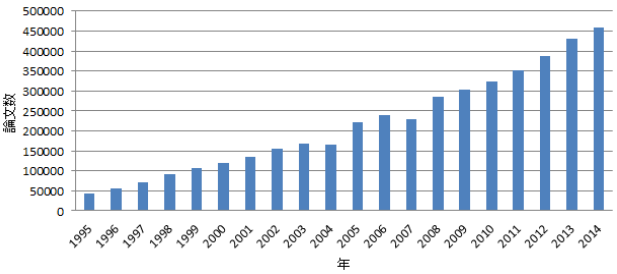

図 5: 各年の被引用論文の数

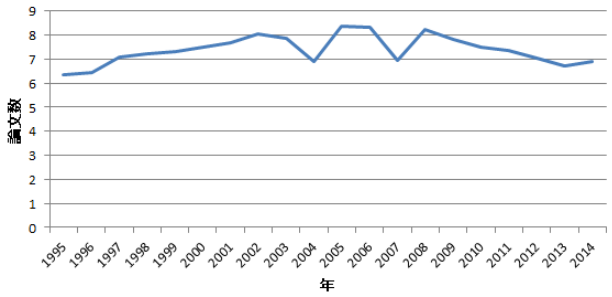

図 6: 各年の論文 1 件あたりの被引用論文数

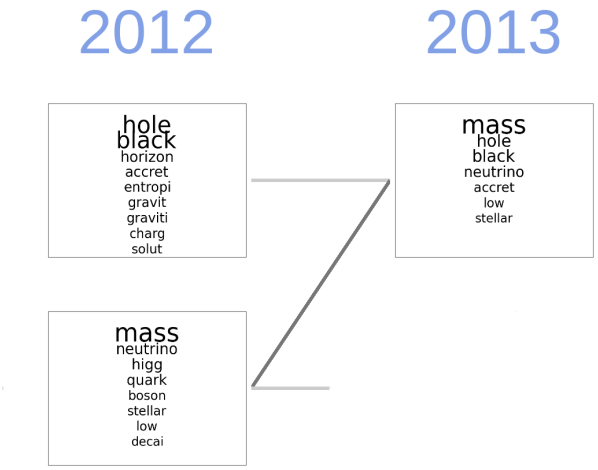

図 8: Jaccard 係数で接続されたトピックの変遷の一部

述べたトピック間の類似度の大きさを表しており，黒い線は同じトピック，灰色の線は似ているトピックを表している．

“black, hole” に関するトピックは 1995 年から 2012 年にかけて少しずつ内容を変えながら発展を続け，2013 年に “mass” に関するトピックと併合，2014 年に再び各々のトピックに分岐している．また，“neutrino” に関するトピックは 2005 年に “mass” に関するトピックと併合し，2013 年に “higgs, boson” に関するトピックに分岐している．以上のことから，本手法によってトピックの変遷の検出，およびトピックの併合，分岐を検出できることが示された．

4.3 コサイン類似度と Jaccard 係数による接続の差異

この節では抽出されたトピックをコサイン類似度と Jaccard 係数でそれぞれ接続し，トピック間の接続の違いに関して議論する．図 8 は Jaccard 係数で検出されたトピックの変遷の一部である．

図 8 における 2012 年のトピック “hole, black, horizon, accret, entropi, gravit, graviti, charg, solut” と 2013 年のトピック “mass, hole, black, neutrino, accret, low, stellar” の接続は，コサイン類似度では検出することができなかった接続である．2013 年の “mass” に関するトピックには “black, hole” という単語が存在しているが，コサイン類似度は接続する際の閾値に達しなかっ

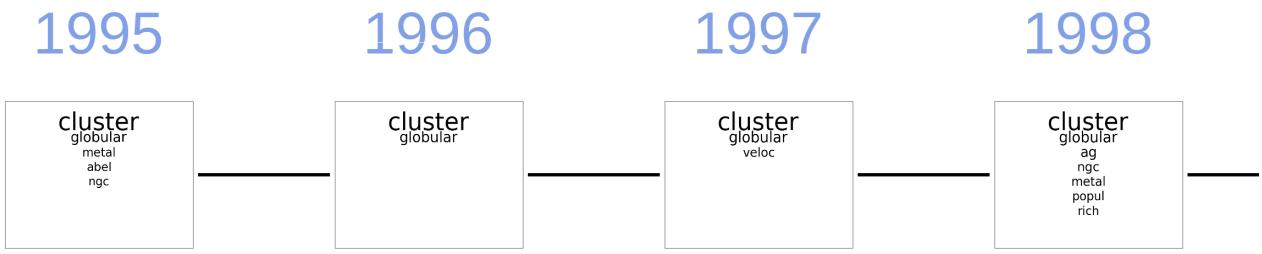

図 9: コサイン類似度で接続されたトピックの変遷の一部

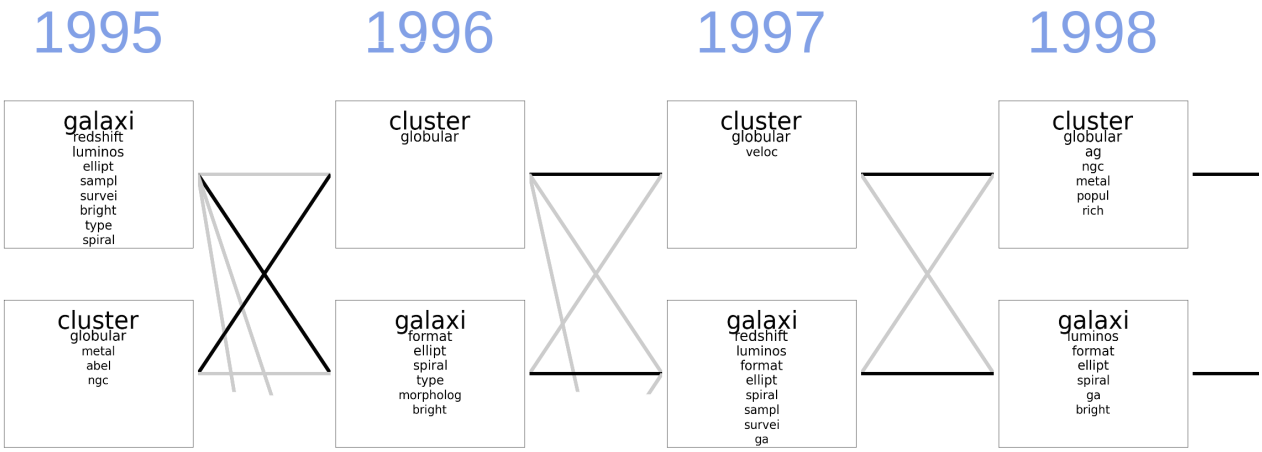

図 10: Jaccard 係数で接続されたトピックの変遷の一部

たため，2012 年の “black, hole” に関するトピックとは接続が行われなかった．

図 9 は “cluster, globular” に関するトピックの変遷をコサイン類似度で接続したもので，図 10 は Jaccard 係数で接続したものである．コサイン類似度による接続では，“cluster, globular” に関する変遷のみ検出された．一方，Jaccard 係数による接続では，“cluster, globular” に関する変遷に加え，“galaxi” に関する変遷との関連を検出することができた．“cluster, globular” は球状星団を意味し，これは “galaxi”，銀河と非常に関連の深いトピックである．

以上のことから，Jaccard 係数によるトピック間の接続は，コサイン類似度では検出できない変遷を検出できることがわかる．

4.4 トピックの変遷の評価

この節では，被引用論文がトピックに与える影響力のパラメータを変化させた際のトピック変遷の評価，および，Jaccard 係数とコサイン類似度によるトピック間の接続の評価に関して議論する．本論文ではトピック変遷を，持続度，多様性という指標で測る．

4.4.1 変遷の持続度

トピックの変遷においては，時間経過によってトピックが出現と消滅を繰り返すよりも，持続的に出現している方が良い．以上の指標をトピック変遷の持続度 *Stability* とし，以下で定式化する．

$$Stability = \sum_{t}\sum_{i}\sum_{j}\sigma_{ij}^{(t)}\cdot\left\{\cos\left(\mathbf{h}_i^{(t)},\mathbf{h}_j^{(t-1)}\right)\right\}^{p} \tag{19}$$

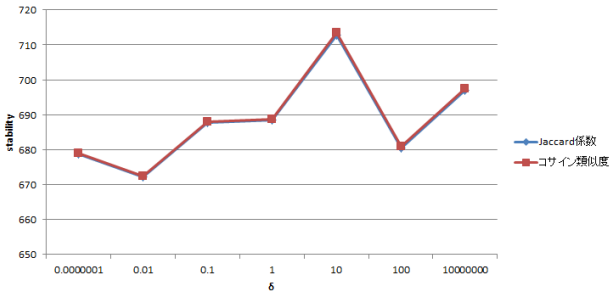

図 11: 接続方法と引用の影響力による *Stability* の差異

ここで，

$$\sigma_{ij}^{(t)} = \begin{cases} 1 & \left(sim\left(\mathbf{h}_i^{(t)},\mathbf{h}_j^{(t-1)}\right) \geq \epsilon_2\right) \\ 0 & \left(sim\left(\mathbf{h}_i^{(t)},\mathbf{h}_j^{(t-1)}\right) < \epsilon_2\right) \end{cases} \tag{20}$$

である．$\cos\left(\mathbf{h}_i^{(t)},\mathbf{h}_j^{(t-1)}\right)$ はトピックを表現する単語の類似度であるため，この値が大きいほど時間経過によるトピックの内容の変遷が小さい．また，時間によるトピックの変遷において，トピックの内容が持続する際には，コサイン類似度が 1 に近くなる傾向にある．したがって，このコサイン類似度の累乗の総和が大きいほど，トピック変遷全体においてトピックの持続性が高いといえる．図 19 における p が大きいほどトピックの持続性の判定が厳しくなり，本論文では検討の結果 $p = 5$ とする．$sim\left(\mathbf{h}_i^{(t)},\mathbf{h}_j^{(t-1)}\right)$ はコサイン類似度による接続のスコア計算の際は $\cos\left(\mathbf{h}_i^{(t)},\mathbf{h}_j^{(t-1)}\right)$，Jaccard 係数による接続のスコア計算の際は $Jaccard\left(\mathbf{h}_i^{(t)},\mathbf{h}_j^{(t-1)}\right)$ とする．図 11 は引用情報がトピックに与える影響力 δ を変化させた際の *Stability* の変化と，コサイン類似度と Jaccard 係数による接続における *Stability* の差異をプロットしたものである．

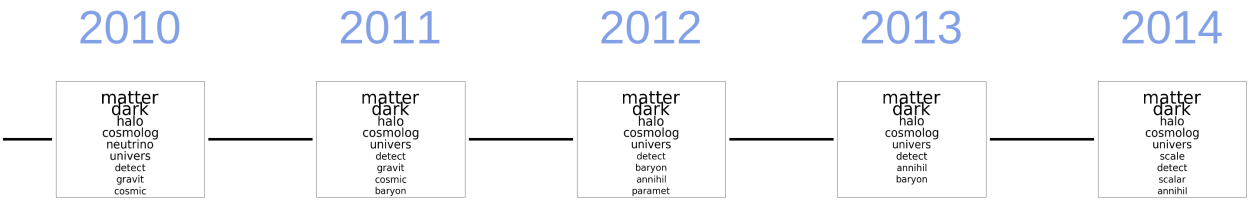

図 12: $\delta = 10$ のときのトピック変遷の一部

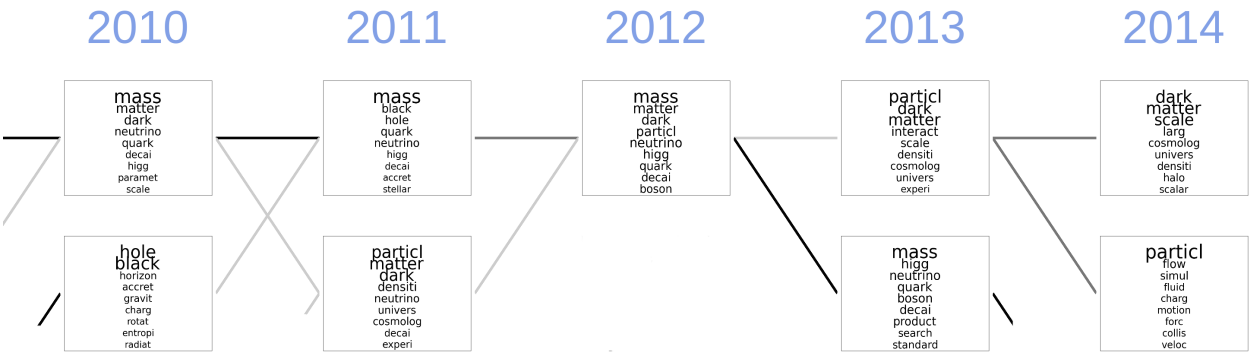

図 13: $\delta = 0.01$ のときのトピック変遷の一部

図 11 を見ると，Jaccard 係数による接続のスコア，コサイン類似度による接続のスコアがほぼ同じである．これは，Jaccard 係数，コサイン類似度による接続で，単語の内容が同じトピックの変遷をほぼもれなく検出できていることを示している．パラメータ δ を変化させた際のスコアの変化を見ると，δ が大きいほどスコアが大きくなる傾向にあることがわかる．これは，引用情報を加えることにより，各時間区分で抽出されるトピックが安定したことを示している．図 12 は *Stability* が一番大きかった $\delta = 10$ でのトピックの変遷の一部で，図 13 は *Stability* が一番小さかった $\delta = 0.01$ でのトピックの変遷の一部である．ここでのトピックの接続は Jaccard 係数を使用している．図 12 を見ると，“dark, matter” に関するトピックが持続して検出されていることがわかる．一方，図 13 を見ると，“dark, matter” に関するトピックが出現，合併，消滅を繰り返している．

4.4.2 変遷の多様性

しかしながら，トピックの持続性のみが変遷の評価指標ではない．持続性とは逆に，どれだけトピックを表現する単語の内容の変化を検出できているかということも変遷の評価指標として考えられる．すなわち，トピックの内容の変化があるほど，そのトピックの変遷の多様性が大きいため，良い変遷といえる．よって変遷の多様性を *Diversity* とし，以下で定式化する．

$$Diversity = -\sum_{t}\sum_{i}\sum_{j}\sigma_{ij}^{(t)}\log_2\left\{\cos\left(\mathbf{h}_i^{(t)},\mathbf{h}_j^{(t-1)}\right)\right\} \qquad (21)$$

$\mathbf{h}_i^{(t)}$ から $\mathbf{h}_j^{(t-1)}$ へトピックが変遷した際の変化の大きさは，$-\log_2\left\{\cos\left(\mathbf{h}_i^{(t)},\mathbf{h}_j^{(t-1)}\right)\right\}$ で表され，その全体の総和はトピック変遷全体における多様性の大きさといえる．図 14 はパラメータ δ を変化させた際の *Diversity* の変化と，コサイン類似度と Jaccard 係数による接続における *Diversity* の差異をプロットしたものである．

図 14 を見ると δ の値によらず，Jaccard 係数による接続のほうが *Diversity* が大きくなっている．これは 4.3 節で議論した，コサイン類似度と Jaccard 係数による接続の差異によるものと考えら

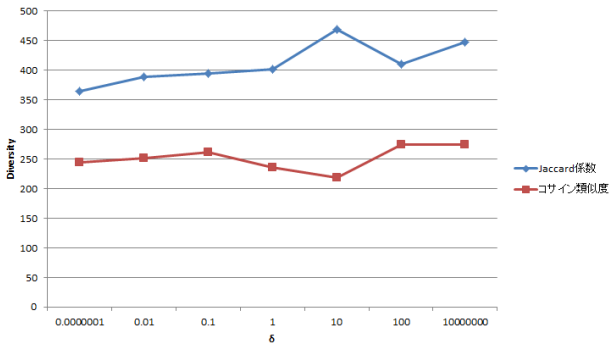

図 14: 接続方法と引用情報の影響力による *Diversity* の差異

れる，コサイン類似度がそれほど高くない場合でも Jaccard 係数による類似度の算出においては接続が行われる場合があるためである．また，δ が大きいほど，*Diversity* が大きくなる傾向にある．

以上の結果から，$\delta = 10$ のときに Jaccard 係数で接続する場合が本手法における一番良いトピック変遷であるといえる．また，トピック変遷の持続性，情報量の観点において，引用情報を組み込むことの有効性，Jaccard 係数で接続することの有効性が実験から示された．

5. 関連研究

トピック抽出，および変遷の検出に関する研究は，現在活発に行われている．トピック抽出，および変遷の検出に関する手法は様々なアプローチが存在しており，確率モデルを用いた手法，グラフ分析を用いた手法，非負値行列分解を用いた手法などが存在する．確率モデルを用いた手法は p-LSI [4] や LDA [2] に代表され，トピックを文書に潜在的に存在している情報として確率的にモデリングし，それによってトピックを抽出する手法である．LDA は様々なかたちで拡張されており，データセットに合わせてトピック数が動的に調節される手法である HDP [14]，引用情報を LDA に組み込むことで論文データベースからトピックの変遷を検出する ITM [3] などがある．

グラフ分析を用いた手法としては [5] などの研究があげられる．この手法は，ある単語を使用している論文同士を引用関係から接続することでグラフを構築する．ある単語がトピックとして振る舞うとき，その単語の引用グラフはよく接続されているはずであるという直感を定式化している．これによりある単語のトピックらしさを算出することができ，2 つの単語からなるトピックを抽出することに成功している．

非負値行列分解を用いた手法は，LDA と比較して計算時間等のコストが軽量であることから，ソーシャルメディアやニュース記事などより即時性が求められるトピック分析によく用いられている．online-NMF [10] は行列分解する際の損失関数に以前観測されたトピックの情報を組み込むことで，twitter などの文書ストリームデータのトピック分析を行うことができる手法である．JPP アルゴリズム [15] は行列分解にトピック間を明示的にリンクさせる行列を組み込み，時間経過によるトピックの変化量に制限を与えることで，ニュースデータにおけるトピックの変遷の検出および妥当性を向上させた手法である．

6. 結論

本論文では，非負値行列分解を用いたトピックの抽出，変遷を検出する手法を提案した．実験より，提案手法によってトピックの抽出および変遷の検出ができることが示された．

今後の課題として，抽出されたトピックの妥当性の検証を行う必要があると考えられる．また別データセットへの適用，異なるパラメータによる比較実験，ユーザインタフェースの拡充などがあげられる．

以上の検証を行い，トピックの変遷の妥当性が十分に証明されたところで，今後の展望として，トピックの変遷を考慮した論文推薦手法の提案等を行っていくことが考えられる．

謝辞

本研究は JSPS 科研費 25330124 の助成を受けたものである．

[文献]

[1] *arXiv*. http://arxiv.org/.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[3] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and C. Lee Giles. Detecting topic evolution in scientific literature: how can citations help? In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China*, pp. 957–966, 2009.

[4] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA*, pp. 50–57, 1999.

[5] 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, San Jose, California, USA*, pp. 370–379, 2007.

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

[7] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. Vol. 401, pp. 788–791. Nature Publishing Group, 1999.

[8] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, Vol. 1. Cambridge university press Cambridge, 2008.

[9] Ramesh Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA*, pp. 542–550, 2008.

[10] Ankan Saha and Vikas Sindhwani. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA*, pp. 693–702, 2012.

[11] Paris Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Independent Component Analysis and Blind Signal Separation*, pp. 494–499. Springer, 2004.

[12] Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng, and Tianyi Wu. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *EDBT 2009, 12th International Conference on Extending Database Technology, Saint Petersburg, Russia*, pp. 565–576, 2009.

[13] Etienne Gael Tajeuna, Mohamed Bouguessa, and Shengrui Wang. Tracking the evolution of community structures in time-evolving social networks. In *2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France*, pp. 1–10, 2015.

[14] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, Vol. 101, No. 476, 2006.

[15] Carmen K. Vaca, Amin Mantrach, Alejandro Jaimes, and Marco Saerens. A time-based collective factorization for topic discovery and monitoring in news. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea*, pp. 527–538, 2014.

[16] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada*, pp. 267–273, 2003.

[17] 澤田宏. 非負値行列因子分解 nmf の基礎とデータ/信号解析への応用. 電子情報通信学会誌, Vol. 95, No. 9, pp. 829–833, 2012.

伊藤 寛祥 Hiroyoshi ITO

筑波大学大学院システム情報工学研究科博士前期課程在学中．情報処理学会，データベース学会学生会員．

天笠 俊之 Toshiyuki AMAGASA

筑波大学システム情報系准教授．データ工学，データベース，Web マイニング等の研究に従事．日本データベース学会，情報処理学会，ACM 各会員．電子情報通信学会，IEEE 各シニア会員．

北川 博之 Hiroyuki KITAGAWA

1978 年東京大学理学部物理学科卒業．1980 年同大学理学系研究科修士課程修了．日本電気（株）勤務の後，1988 年筑波大学電子・情報工学系講師．同助教授を経て，現在，筑波大学システム情報系教授，ならびに計算科学研究センター教授．理学博士（東京大学）．データベース，情報統合，データマイニング，情報検索等の研究に従事．日本データベース学会会長，情報処理学会フェロー，電子情報通信学会フェロー，ACM，IEEE，日本ソフトウェア科学会，各会員．