5N-03

# 非負値行列分解を用いた 論文データベースにおけるトピックの変遷の検出

伊藤　寛祥†　　天笠　俊之‡　　北川　博之‡

†筑波大学情報学群情報科学類　　‡筑波大学システム情報系

## 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。著名な例としては DBLP[1]、ADS[2]、ArXiv[3]、PUPMED[4]、MEDLINE[5]などがある。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、研究者ネットワークの抽出などが可能となる。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[8]。

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[6]、LDA[7]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非負行列分解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。

そこで、本研究では非負行列分解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、非負行列分解を適用し、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、共通する文章を手掛かりに関連づけ、トピックの変遷を検出する。

## 2 非負値行列分解（NMF）

NMF（Non-negative Matrix Factorization）とは入力として与えられた非負行列を近似的に二つの非負行列に分解する手法である。NMF を使用すると、入力データ行列における頻出パターンを抽出できる。NMF は画像認識における特徴量抽出や信号処理におけるスペクトル分解などにも使用されている[9]。

文書をベクトル化して得られる、$N_d \times N_f$ の行列$A$（$N_d$ は文書数、$N_f$ は単語数）に対して NMF を適用すると、$A \approx WH$ を満たす、$K \times N_f$ の行列 $H$ と、$N_d \times K$ の行列 $W$ の二つの行列に分解できる（$K$ はトピック数）。

ここで、行列 $H$ を見ることで、各トピックの単語分布がわかる。また行列 $W$ を見ることで、各文書がどのトピックにどれだけ関連しているかを知ることができる。それによってトピックに基づく文書のソフトクラスタリングが可能となる。

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 $W \geq 0, H \geq 0$ という条件の下で最適化することで行列分解ができる。

$$L = \underset{W,H}{arg\,min} \|A - WH\|_F^2 \quad \cdots (1)$$

$\|\cdot\|_F^2$：フロベニウスノルム

## 3 提案手法

本論文では NMF をベースにおいた、トピックの変遷を検出する手法を提案する。

### 3.1　文書集合の行列化

提案手法では、時間経過によるトピックの変遷の検出するため、データセット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、隣接する時間区分同士をオーバーラップさせる。

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 $t$ における論文を並べて行列としたものを $X^{(t)}$ 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掛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 $X^{(t)}$ が引用している論文集合を $X^{(t)}$ 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の数）。

### 3.2　トピックの抽出

この節では引用情報を組み込んだ、トピック抽出のための新しい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

---

*Detecting topic evolution from scientific bibliography database using non-negative matrix factorization*
Hiroyoshi Ito† (hiro.3188@kde.cs.tsukuba.ac.jp), Toshiyuki Amagasa‡ (amagasa@cs.tsukuba.ac.jp) and Hiroyuki Kitagawa‡ (kitagawa@cs.tsukuba.ac.jp)
†College of Information Science, University of Tsukuba
‡Faculty of Engineering,Information and Systems, University of Tsukuba

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 1）。行列分解は式 (2) に示す損失関数の最適化によって行う。

$$L = \underset{W_X^{(t)}, W_C^{(t)}, H^{(t)}}{arg\ min} \left\| X^{(t)} - W_X^{(t)} H^{(t)} \right\|_F^2 + \delta \left\| C^{(t)} - W_C^{(t)} H^{(t)} \right\|_F^2 + \alpha \left\| H^{(t)} \right\|_1 + \beta \left\| \begin{pmatrix} W_X^{(t)} \\ W_C^{(t)} \end{pmatrix} \right\|_1 \quad \cdots (2)$$

パラメータ $\delta$ によって引用文献がトピックに与える影響力を調節できる。パラメータ $\delta$ が 0 に近づくほど引用論文がトピックに与える影響力は小さくなり、$\infty$ に近づくほど引用論文がトピックに与える影響力は大きくなる。

損失関数は KKT 条件より導出される更新式 (3)、(4)、(5) を適用し最適化する。

$$W_X^{(t)} = W_X^{(t)} * \frac{\left[ X^{(t)} H^{(t)^T} - \beta \right]}{\left[ W_X^{(t)} H^{(t)} H^{(t)^T} \right]} \quad \cdots (3)$$

$$W_c^{(t)} = W_c^{(t)} * \frac{\left[ C^{(t)} H^{(t)^T} - \beta \right]}{\left[ W_C^{(t)} H^{(t)} H^{(t)^T} \right]} \quad \cdots (4)$$

$$H^{(t)} = H^{(t)} * \frac{\left[ W_X^{(t)^T} X^{(t)} + \delta W_C^{(t)^T} C^{(t)} - \alpha \right]}{\left[ W_X^{(t)^T} W_X^{(t)} H^{(t)} + \delta W_c^{(t)^T} W_C^{(t)} H^{(t)} \right]} \quad \cdots (5)$$

$*$ ：アダマール積

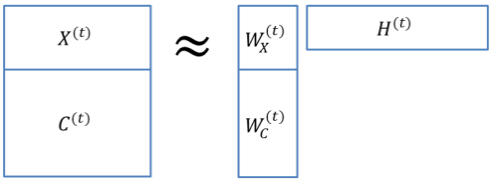


図 1. 引用論文を考慮した行列分解

出力された行列 $W_X^{(t)}$ は各文書がも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。時刻 $t$ 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文書集合を $T_i^{(t)}$ であらわす。

### 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。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の文書数を利用することが考えられる。本研究では検討の結果後者を利用することとした。具体的には以下のように定式化される。

$$sim\left(h_i^{(t)}, h_j^{(t-1)}\right) = \frac{\left| T_i^{(t)} \cap T_j^{(t-1)} \right|}{\left| T_i^{(t)} \cup T_j^{(t-1)} \right|} \quad \cdots (6)$$

・$h_i^{(t)}$:　時間 $t$ における $i$ 番目のトピック

## 4. 実験

提案手法の評価のために、論文データベース ArXiv[3] を対象に 1995 年から 2014 年までの論文、およびそれぞれの論文が引用している論文を取得し、手法を適用した。

図 4 は本手法で検出されたトピックの変遷の一部である。図における四角形は抽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。2012 年におけるトピック「hole, black, horizon, accret, entropi, gravit, charg, solut」を起点として、2011 年、2013 年のトピックの中から類似度の大きいものを探して結合させたものであり、線の濃さはトピック間のつながりの強さを表している。トピックの内容の変遷、およびトピックの分岐、併合を検出できていることがわかる。

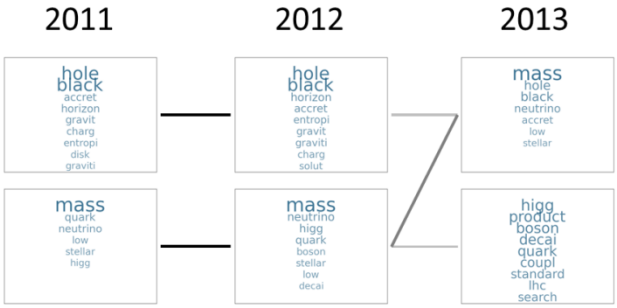


図 4. トピックの変遷の例

## 5. まとめ

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出来ることが示された。

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挙げられる。

## 謝辞

本研究は JSPS 科研費 25330124 の助成を受けたものです。

## 参考文献

[1] DBLP : http://www.informatik.uni-trier.de/~ley/db/.
[2] ADS : http://adsabs.harvard.edu/index.html
[3] ArXiv : http://xxx.yukawa.kyoto-u.ac.jp/
[4] PUBMED : http://www.ncbi.nlm.nih.gov/pubmed
[5] MEDLINE : http://www.ebsco.co.jp/medical/medline/
[6] T.Hofmann et al, “Probabilistic Latent Semantic Indexing”, Proceedings of the Twenty-Second Annual International SIGIR Conference 2003
[7] D.Blei et al, “Latent Dirichlet allocation”, Journal of Machine Learning Research 2003.
[8] Q.He et al, “Detecting Topic Evolution in Scientific Literature: How Can Citations Help?”,CIKM’09
[9] 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.