社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－NL－179 (2)
2007－SLP－ 66 (2)
2007／5／24

NMF とリンクベースの修正法によるピンポン型文書クラスタリング

新納浩幸　　佐々木稔

茨城大学工学部情報工学科
{shinnou,msasaki}@mx.ibaraki.ac.jp

Non-negative Matrix Factorization (NMF) は効果的な文書クラスタリング手法である。本論文では NMF の精度をさらに高めるために、NMF とリンクベースの修正法を交互に適用するピンポン型文書クラスタリング手法を提案する。NMF をピンポン型で利用することで、効果的な文書クラスタリングが期待できるが、NMF は入力のクラスタリング結果を改善できない場合も多く、ピンポン型で利用すると悪影響も多い。ここでは、ピンポンの終了条件の設定でこの問題に対処する。具体的には、リンクベースの修正法の適用の後に、設定した評価関数の値が改善されたかどうかでピンポンの終了を判定する。もし改善されていなければ、ピンポンを終了し、前回のリンクベースの修正法を行った後の結果を最終的なクラスタリング結果とする。これによって、NMF を利用した効果的なピンポン型クラスタリングが可能となる。実験では 16 個の文書データセットを利用して、本手法を k-means や NMF と比較した。基本となる NMF の結果を大きく改善できた。

Ping-Pong Document Clustering by using NMF and Linkage Based Refinement

Hiroyuki Shinnou　　Minoru Sasaki
Department of Computer and Information Sciences, Ibaraki University
{shinnou,msasaki}@mx.ibaraki.ac.jp

Non-negative Matrix Factorization (NMF) is a powerful document clustering method. This paper proposes a ping-pong document clustering method using NMF and the linkage based refinement alternately, in order to improve the clustering result of NMF. The use of NMF in the ping-pong strategy can be expected effective for document clustering. However, NMF in the ping-pong strategy often worsens performance because NMF often fails to improve the clustering result given as the initial values. Our method handles this problem with the stop condition of the ping-pong process. Concretely speaking, our method decides the stop/continue of the ping-pong process by the value of an object function for the clustering result produced by the linkage based refinement. If that value is not improved, our method stops the ping-pong process, and outputs the clustering result produced by the linkage based refinement in the previous ping-pong process. By this setting, our method can use NMF in the ping-pong strategy. In the experiment, we compared our method with the k-means and NMF by using 16 document data sets. Our method improved the clustering result of NMF significantly.

1 はじめに

Non-negative Matrix Factorization (NMF) は次元縮約を応用したクラスタリング手法であり、文書クラスタリングのようにデータが高次元かつスパースとなる場合に効果がある。本論文では NMF から得られるクラスタリング結果の精度をさらに高めるために、NMF とリンクベースの修正法を交互に適用するピンポン型文書クラスタリング手法を提案する。

ピンポン型クラスタリングでは、クラスタリング結果を改善する2つの手法を用意し、それら手法を交互に利用することで、クラスタリング結果の精度を高めてゆく。「ピンポン型」という用語は一般的ではないが、論文 [1] ではこの手法のことを "ping-pong strategy" と名付けていることから本論文ではこの用語を用いることにした。ピンポン型クラスタリングで利用される2つの手法は、どちらもランダムなクラスタリング結果の初期値を与えることで、単独のクラスタリング手法としても機能するが、ピンポン型クラスタリングでは、それらを単独に利用したクラスタリングよりも、精度の高いクラスタリング結果を得ることができる。

ピンポン型クラスタリングの代表的研究は、Dhillon らの local search である [1]。そこでは k-means とクラスタリング結果を修正する first valiation という手法を組み合わせている。また Ding らは、NMF と pLSI が本質的に同じ評価関数を使っているが、その最適解への探索手法が異なることを利用して、それらを組み合わせたピンポン型クラスタリングを提案している [3]。本論文では2つの手法として NMF とリンクベースの修正法を用いる。以下、本論文ではリンクベースの修正法（Linkage Based Refinement）を LBR と略す。

NMF は次元縮約の手法を応用したクラスタリング手法である [5]。今、クラスタリング対象の m 次元で表現された n 個の文書を m 行 n 列の索引語文書行列 X で表す。目的とするクラスタの数が k である場合、NMF では X を以下のような行列 U と V^T に分解する。

$$X = UV^T$$

ここで U は m 行 k 列、V は n 行 k 列である。V^T は V の転置を表す。また U と V の要素は非負である。行列 V が文書ベクトルを次元縮約した結果であり、縮約されたベクトルを利用してクラスタリングを行ってもよいが、NMF では行列 V 自体がクラスタリング結果を表している。

行列 V と U は、それらの初期値 V_0 と U_0 を用いて、ある単純な繰り返し処理から得られる [4]。ここで V_0 はあるクラスタリング結果に対応するので、NMF は初期値のクラスタリング結果を改善する手法とも捉えることができる。そのためピンポン型クラスタリング手法の構成手法として利用できる。また NMF は文書クラスタリングに対して効果的な手法なので、ピンポン型の構成手法として利用することで、より効果的な文書クラスタリングが期待できる。

LBR はグラフスペクトル理論を用いたクラスタリング手法である Mcut の結果を修正する目的で提案されたものである [2]。この修正法では、クラスタに属するデータを別クラスタに移動させた場合の評価関数を設定する。各データに対するクラスタをその評価関数によって再設定することで修正を行う。この修正法はヒューリスティクスであり、クラスタリングの評価関数の値を改善する保証はないが、グラフスペクトル理論によるクラスタリング結果に対して有効であることが実験により示されている [2]。ここではこの修正法が、一般のクラスタリング結果に対しても有効であると考え、ピンポン型クラスタリングの手法として利用する。

本手法の特徴はピンポン型クラスタリングの構成手法として NMF を採用している点である。NMF は文書クラスタリングに対して効果的な手法であり、ピンポン型にすることにより文書クラスタリングに対して更なる精度向上が期待できる。しかし NMF は入力のクラスタリング結果を改善できる場合は少なく、ピンポン型に利用すると悪影響も多い。ここでは、ピンポンの終了条件の設定でこの問題に対処する。具体的には、LBR の適用の後に、Mcut の評価関数の値が改善されたかどうかでピンポンの終了を判定する。もし改善されていなければ、ピンポンを終了し、前回の LBR を行った後の結果を最終的なクラスタリング結果とする。

実験では、16 個の文書データセットに対して、本手法の他、クラスタリングの標準手法である k-means、NMF 及び NMF の結果を LBR で修正する手法によりクラスタリングを行う。エントロピーで評価を行い、本手法の有効性を示す。

2 NMF

NMF は $m \times n$ の索引語文書行列 X を、$m \times k$ の行列 U と $n \times k$ の行列 V の転置行列 V^T の積に分解する [5]。ここで k はクラスタ数である。

$$X = UV^T$$

NMF はクラスタに対応したトピックの次元を k 個想定し、その基底ベクトルの線形和によって、文書ベクトル及び索引語ベクトルを表現することに対応する。つまり基底ベクトルの係数が、そのトピックとの関連度を表しているので、行列 V 自体がクラスタリング結果と見なせる。具体的には、i 番目の文書 d_i は、行列 X の第 i 列のベクトルで表現され、その次元縮約された結果が、行列 V の第 i 行のベクトルとなる。このとき、V の第 i 行のベクト

ルは

$$(v_{i1}, v_{i2}, \cdots, v_{ik})$$

と表せ、文書 d_i のクラスタは

$$\arg\max_{j \in 1:k} v_{ij}$$

となる。

与えられた索引語文書行列 X から、U と V は以下の繰り返しで得ることができる [4]。

$$u_{ij} \leftarrow u_{ij} \frac{(XV)_{ij}}{(UV^TV)_{ij}} \tag{1}$$

$$v_{ij} \leftarrow v_{ij} \frac{(X^TU)_{ij}}{(VU^TU)_{ij}} \tag{2}$$

ここで u_{ij} と v_{ij} はそれぞれ U と V の i 行 j 列の要素を表す。また $(X)_{ij}$ により行列 X の i 行 j 列の要素を表す。

また各繰り返しの後に U を以下のように正規化する。

$$u_{ij} \leftarrow \frac{u_{ij}}{\sqrt{\sum_i u_{ij}^2}} \tag{3}$$

繰り返しの終了は、繰り返しの最大回数を決めておくか、UV^T と X との距離 J の変化量から判定する。

$$J = ||X - UV^T||_F \tag{4}$$

通常、行列 V と U の初期値 V_0 と U_0 は、ランダムな値を与えることで作成される。ただし式 1 と 2 による繰り返しは局所最適解にしか収束しないために、V_0 と U_0 によって、最終的に得られる V と U は大きく異なり、結果としてクラスタリングの精度も大きく異なる。

一方、行列 V はクラスタリング結果を表していることを考えれば、NMF は初期値のクラスタリング結果を改善する手法と捉えることもできる。そのためにより精度の高い初期値を与えることで、より精度の高いクラスタリング結果を得ることが期待できる。

3 LBR

本論文で使用するもうひとつのクラスタリング結果を改善する手法は、論文 [2] で提案された LBR である。

LBR はグラフスペクトル理論を用いたクラスタリング手法である Mcut の結果を修正する目的で提案された。グラフスペクトル理論を用いたクラスタリング手法では、その結果に「ねじれ現象」と呼ばれる不具合が生じることがあり、それを解消するために提案された手法が LBR である。

まず Mcut の概略を述べる。Mcut ではデータをグラフのノードとして表現し、ノード間のエッジの重みには両端のデータ間の類似度を与える。類似度が 0 の場合は、エッジを張らない。このようにデータの集合をグラフとして表した場合、クラスタリングとはエッジをカットして、全体のグラフをいくつかのサブグラフに分割することに対応する。その際に、サブグラフ内のエッジは密になり、サブグラフ間でカットしたエッジは疎になるようなカットが望ましい。望ましいカットを見つけるために評価関数を設定する。

サブグラフ A と B の類似度 $cut(A, B)$ を以下で定義する。

$$cut(A, B) = W(A, B) \tag{5}$$

ここで関数 $W(A, B)$ はサブグラフ A と B の間にあるエッジの重みの総和である。また、$W(A) = W(A, A)$ と定義する。

Mcut の評価関数は以下である。

$$Mcut = \frac{cut(A, B)}{W(A)} + \frac{cut(A, B)}{W(B)} \tag{6}$$

式 6 を最小化するようなサブグラフ A と B を見つけることが課題であるが、式 6 の最小化の問題は、ある固有値問題を解くことで、その近似解を得ることができる。「ねじれ現象」とはこの近似解に対する不具合である。

Mcut は 2 つのクラスタに分割するのが基本である。目的のクラスタ数を得るまで、上記の処理を再帰的に繰り返す。

LBR とは、クラスタ A に属するデータ u をクラスタ B に移動させたときに生じる得点をある評価関数で算出し、その値が正であるときに u のクラスタを B に移動させる。その評価関数は以下である。

$$\Delta l_{AB}(u) = l(u, B) - l(u, A)$$

ここで関数 $l(u, X)$ の定義は以下である。

$$l(u, X) = \frac{1}{|X|} \sum_{v \in X} sim(u, v)$$

また $sim(u, v)$ は u と v の類似度を表す。$\Delta l_{AB}(u) < 0$の場合、クラスタは変更されない。

上記が LBR の基本的な考え方である。Mcut は 2 分割を再帰的に繰り返して、一般の分割数まで分割するので、各 2 分割の後に上記の修正法を行えばよい。

一般のクラスタ数が k であるクラスタリング結果 $\{G_1, G_2, \cdots, G_k\}$ に対する LBR を説明する。

まずクラスタリング結果 $\{G_1, G_2, \cdots, G_k\}$ に対する Mcut の評価関数は以下である。

$$Mcut_K = \frac{cut(G_1, G_1)}{W(G_1)} + \frac{cut(G_2, \bar{G_2})}{W(G_2)} + \cdots + \frac{cut(G_k, \bar{G_k})}{W(G_k)} \tag{7}$$

この評価関数は値が低いほどよいことを意味する。また $\bar{G}_i$ は集合 G_i の補集合を表す。

今、データ u が G_i の要素である場合、$\Delta l_{ij}(u)$ を以下で定義する。

$$\Delta l_{ij}(u) = l(u, G_j) - l(u, G_i)$$

そして、

$$\hat{i} = \arg\max_j \Delta l_{ij}(u)$$

とおき、$i \neq \hat{i}$ の場合、クラスタ G_i にあったデータ u をクラスタ $G_{\hat{i}}$ へ移動させる。

全てのデータ u に対して、上記の処理が終わり、新たなクラスタリング結果 $\{G_1, G_2, \cdots, G_k\}$ が得られた後に、再び上記の処理を繰り返す。データの移動が起こらなくなるまで、上記の処理を繰り返す。これが LBR である。

注意として、LBR を行った場合、必ずしも評価関数（式 7）の値が改善されるということはない。つまり LBR は、ヒューリスティクスな修正法となっている。

4 ピンポン型クラスタリング

本手法で提案するピンポン型クラスタリングではまず NMF により、あるクラスタリング結果を導く。次に LBR によりそのクラスタリング結果を改善する。改善されたクラスタリング結果を利用して、NMF の初期値となる行列 V_0 と U_0 を作成し、NMF を実行する。この処理を繰り返す。

V_0 と U_0 の作成法を述べる。クラスタリング結果によりデータ i がクラスタ c となっていた場合、以下によって V_0 の第 i 行のベクトルを構築する。また U_0 は XV_0 によって構築する。

$$v_{ij} = \begin{cases} 1.0 & (j = c) \\ 0.1 & (j \neq c) \end{cases}$$

NMF も LBR もどちらも入力となるクラスタリング結果を改善することが理想であるが、その保証はない。特に NMF は現実には入力なるクラスタリング結果を改善できない場合も多く、ピンポン型の構成手法として利用することは困難である。

ここでは、ピンポンの終了条件の設定によりこの問題に対処する。まず LBR を行った後に、評価関数である式 7 の値を計算する。前回の LBR を行った後の式 7 の値よりも改善されていれば、処理を続行し、改善されていなければ、前回の LBR を行った後の結果を最終的なクラスタリング結果とする。

例を示す。図 1 は実験で用いた文書データセット tr12 に対する本手法のピンポン型クラスタリングを行った結果を示している。縦軸は評価関数（式 7）の値を表す。

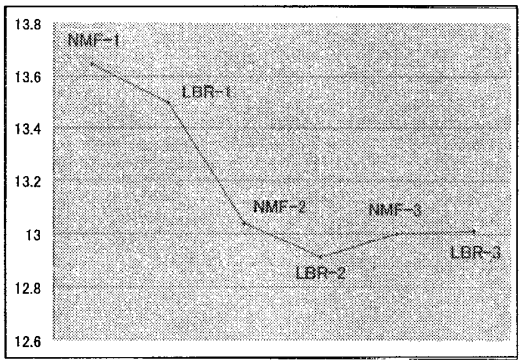

図 1: ピンポンによる評価関数の値の変化（1）

まず NMF によりクラスタリングを行う。その結果に対する評価関数の値が図 1 の NMF-1 である。次に、そのクラスタリング結果に対して LBR を行う。その結果に対する評価関数の値が LBR-1 である。次にそのクラスタリング結果を元に NMF の初期値を作成し、NMF によりクラスタリングを行う。その結果に対する評価関数の値が NMF-2 である。次に、そのクラスタリング結果に対して LBR を行う。その結果に対する評価関数の値が LBR-2 である。LBR-2 と LBR-1 の値を比較し、LBR-2 の値の方が小さいので、ピンポンを続行する。上記の処理を繰り返し、LBR-3 を得る。ここで LBR-3 と LBR-2 の値を比較すると、LBR-3 の値の方が大きいので、ここでピンポンを終了させ、LBR-2 に対するクラスタリング結果を出力とする。

上記の例は NMF も LBR もどちらも入力となるクラスタリング結果を、評価関数の上では、改善している。この場合、LBR-2 から NMF-3 に移る時点で評価関数の値が悪くなるので、LBR-2 でピンポンを終了させればよく、実際に LBR-3 を計算するのは無駄であった。

しかし多くのデータセットに対して NMF では、クラスタリング結果を、実際にもそして評価関数の上でも、改善できない場合が多い。例えば、図 2 は実験で用いた文書データセット kb1 に対する本手法のピンポン型クラスタリングを行った結果を示している。

図 2 においては、LBR-3 の時点でピンポンが終了し、LBR-2 のクラスタリング結果が出力される。図から分かるように、NMF-2 は LBR による結果（LBR-1）を評価関数の上では改善できていない。しかし、NMF-2 は NMF-1 よりも改善されており、しかも NMF-2 を LBR で修正した結果の LBR-2 は LBR-1 よりも評価関数の値が小さい。つまり単純に NMF による評価関数の悪化によりピンポンを終了させるのは得策ではない。本手法によるピンポンの終了条件は図 2 のような変化のパターンを考慮して考案したものである。

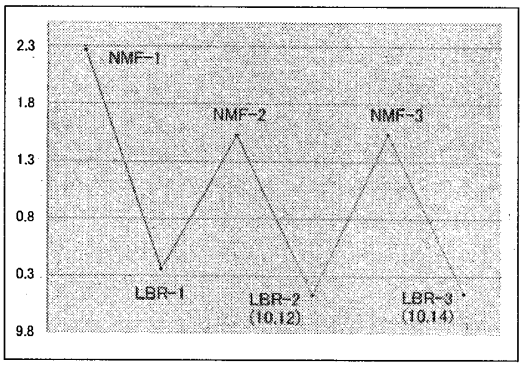

図 2: ピンポンによる評価関数の値の変化（2）

5 実験

ここでの実験では以下のサイトで公開されている文書データセットのうち表 1 の 16 文書データセットを用いる。各データにおける次元の値は正規化されていないので、ここでは TF-IDF によって正規化を行った。

http://glaros.dtc.umn.edu/gkhome/
cluto/cluto/download

表 1: 文書データセット

Data	# of documents	# of terms	# of classes
cranmed	2431	41681	2
fbis	2463	2000	17
hitech	2301	126373	6
k1a	2340	21839	20
k1b	2340	21839	6
la1	3204	31472	6
la2	3075	31472	6
re0	1504	2886	13
re1	1657	3758	25
reviews	4069	126373	5
tr12	313	5804	8
tr23	204	5832	6
tr31	927	10128	7
tr41	878	7454	10
tr45	690	8261	10
wap	1560	6460	20

用いたクラスタリング手法は、クラスタリングの標準手法である k-means、NMF、NMF の結果を LBR で修正するもの（NMF+LBR）、および本手法（Ping-Pong）の 4 つである。Ping-Pong と NMF+LBR の違いは、クラスタリング結果のやりとりを繰り返すかどうかである。NMF+LBR は LBR で修正した結果を NMF に返さずにそのまま出力する。一方、Ping-Pong はその結果を NMF に返し、クラスタリング結果のやりとりを繰り返す。結果を表 2 に示す。“KM”、“NMF” が k-means 及び NMF の結果を表し、“NMF+LBR” が NMF の結果を LBR で修正したものを表し、“PP(NMF)” が本手法のピンポン型クラスタリングの結果を示す。

また表の値はエントロピーである。エントロピーはクラスタリング結果を評価するための 1 つの尺度である。データセットのクラスタリングの正解が $\{K_h\}_{h=1}^{k}$ であり、得られたクラスタリングが $\{C_j\}_{j=1}^{k}$ であるとき、クラスタ C_i に対するエントロピーは以下で定義される。

$$E_i = -\sum_{h=1}^{k} P(K_h|C_i) \log P(K_h|C_i)$$

各クラスタに対して E_i を求め、クラスタのデータ数による重み付き平均をとることで全体のエントロピーが定義される。つまり、エントロピーの値が小さいほどクラスタリング結果が良好であることを意味する。

表 2 から本手法の有効性が確認できる。

表 2: クラスタリング結果

Data	KM	NMF	NMF+LBR	PP(NMF)
cranmed	0.106	0.748	0.067	**0.055**
fbis	**0.330**	0.383	0.360	0.358
hitech	**0.597**	0.724	0.678	0.679
k1a	0.403	0.384	0.370	**0.352**
k1b	0.306	0.277	0.233	**0.218**
la1	0.660	0.547	0.430	**0.401**
la2	0.620	0.565	**0.411**	0.413
re0	0.384	0.397	**0.373**	0.386
re1	0.391	0.355	**0.310**	0.316
reviews	0.406	0.602	**0.323**	**0.323**
tr12	0.641	0.424	0.406	**0.357**
tr23	0.484	0.473	**0.382**	0.399
tr31	0.373	0.393	0.327	**0.310**
tr41	0.381	0.269	0.277	**0.242**
tr45	0.473	0.254	**0.210**	0.247
wap	0.427	0.378	0.378	**0.371**
平均	0.436	0.448	0.346	**0.339**

6 考察

ピンポン型クラスタリングでは、利用する 2 つの手法が以下の条件を満たしていることが必要である。

条件 1 入力はクラスタリング結果であること

条件 2 出力は入力のクラスタリング結果を改善すること

本論文で提案するピンポン型クラスタリング手法では、2 つの手法として NMF と LBR を利用した。どちらの手法も条件 1 は満たしているが、条件 2 を満たしている保証はない。

本実験の NMF と NMF+LBR を比較すると、16 個中 14 個でエントロピーが減少している。残り 2 つのうち、1 つは変化なしであり（wap）、エントロピーが増加しているのは 16 個中 1 個である（tr41）。

この結果から、LBR は条件 2 をほぼ満たしていると考えることができる。

次に、先の実験で、最初に LBR から返されたクラスタリング結果を NMF が改善したかどうかを調べたところ、16 個中 5 個でエントロピーが減少したが、残り 11 個ではエントロピーが増加している。この結果から、NMF は条件 2 を満たさないことも多いことが分かる。

この問題は NMF の評価関数（式 4）に由来する。NMF のアルゴリズムを用いれば、式 4 の値は改善されてゆくが、それがクラスタリング結果の精度を改善してゆくことに繋がってはいない。この問題のために、ピンポンで NMF にクラスタリング結果を渡しても、その結果が改善されず、逆に改悪されて戻ってきてしまう場合がある。この問題は論文 [6] でも指摘されている。そこでは NMF のアルゴリズムの終了条件を別の評価関数を利用することで、この問題に対処している。

ただし本手法では LBR の結果だけを使って、ピンポン処理の終了を判定し、最終的なクラスタリング結果は LBR によるものを出力しているので、仮に NMF が入力のクラスタリング結果を改善できなくとも、悪影響は少ない。

NMF の入力となるクラスタリング結果の精度と出力のクラスタリング結果の精度との関係を調べることは今後の課題である。

またピンポン型クラスタリングで利用できる他手法として k-means は典型的である。参考として、k-means と LBR を利用したピンポン型クラスタリングを行った結果を表 3 に示す。“KM+LBR” が k-means の結果を LBR で修正した結果を示し、“PP(KM)” が k-means と LBR によるピンポン型クラスタリングの結果を示す。

表 3: k-means を利用したピンポン型クラスタリングの結果

Data	KM	KM+LBR	PP(KM)	PP(NMF)
cranmed	0.106	0.070	0.070	**0.055**
fbis	0.330	**0.325**	**0.325**	0.358
hitech	**0.597**	0.619	0.613	0.679
k1a	0.403	0.387	0.376	**0.352**
k1b	0.306	0.246	0.240	**0.218**
la1	0.660	0.440	0.425	**0.401**
la2	0.620	0.421	0.421	**0.413**
re0	0.384	0.385	**0.379**	0.386
re1	0.391	0.351	0.330	**0.316**
reviews	0.406	0.358	0.364	**0.323**
tr12	0.641	0.422	**0.321**	0.357
tr23	0.484	0.457	0.457	**0.399**
tr31	0.373	**0.235**	**0.235**	0.310
tr41	0.381	0.312	0.318	**0.242**
tr45	0.473	**0.195**	0.261	0.247
wap	0.427	0.378	**0.361**	0.371
平均	0.436	0.350	0.343	**0.339**

この結果からも、LBR は条件 2 をほぼ満たしていることが確認できる。またピンポン型クラスタリングの手法として NMF と k-means の優劣は微妙であった。上記実験では NMF が若干良いが、その差はわずかである。

ただし NMF は k-means よりも、出力結果がより多くの情報を含んでいる。例えばデータがクラスに属する度合い、単語がクラスに関連する度合いなどの情報が分解後の行列に記載されている。このために更に何らかの修正を加味させてゆく場合には、NMF の方が発展性があると考える。

NMF の初期値と出力結果の精度の関係、および NMF の出力に含まれるその他の情報を修正に利用する手法の考案を今後の課題とする。

7 おわりに

本論文は NMF の精度をさらに高めるために、NMF と LBR を交互に適用するピンポン型文書クラスタリング手法を提案した。NMF も LBR も入力となるクラスタリング結果を改善する保証はないが、実験では、LBR は 16 個中 14 個で改善し、多くの場合、クラスタリング結果を改善できることを示した。NMF は 16 個中 5 個を改善したにとどまるが、ピンポンの終了条件を提案したように LBR にまかせることで、ピンポン型クラスタリングの構成要素として利用できる。また提案手法により最終的に得られたクラスタリング結果は、基本となる NMF の結果を大きく改善できた。NMF の初期値と出力結果の精度の関係、および NMF の出力に含まれるその他の情報を修正に利用する手法の考案を今後の課題とする。

参考文献

[1] Inderjit S. Dhillon, Yuqiang Guan, and J. Kogan. Iterative Clustering of High Dimentional Text Data Augmented by Local Search. In *The 2002 IEEE International Conference on Data Mining*, pp. 131–138, 2002.

[2] Chris Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, and Horst Simon. Spectral Min-max Cut for Graph Partitioning and Data Clustering. In *Lawrence Berkeley National Lab. Tech. report 47848*, 2001.

[3] Chris Ding, Tao Li, and Wei Peng. Nonnegative Matrix Factorization and Probabilistic Latent Semantic Indexing: Equivalence, Chi-square Statistic, and a Hybrid Method. In *AAAI National Conf. on Artificial Intelligence (AAAI-06)*, 2006.

[4] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pp. 556–562, 2000.

[5] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of SIGIR-03*, pp. 267–273, 2003.

[6] 新納浩幸, 佐々木稔. Mcut+NMF による文書クラスタリング. 言語処理学会第 13 回年次大会, pp. 558–561, 2007.