Mcut + NMF による文書クラスタリング

新納浩幸
茨城大学工学部情報工学科

佐々木稔
茨城大学工学部情報工学科

1 はじめに

文書クラスタリングとは文書の集合をトピックなどの観点からいくつかのグループに分割するタスクである．文書クラスタリングは文書の集合に対して，知的な処理を行う基本的な処理であり，テキストマイニングの重要な構成要素となっている [13]．具体的な応用としては，検索文書を絞り込むために，検索された文書の集合をクラスタリングする適合性フィードバックが盛んに研究されている ([11] など)．

文書クラスタリングでは，まずデータとなる文書をベクトルで表現する．データがベクトルで表現できれば，ward 法や k-means などの古典的クラスタリング手法を用いることができる．文書をベクトルで表現するためには，通常，bag of words のモデルを用い，次に TF-IDF などによって次元の重みを調整する．このようにして作成されたベクトルは高次元かつスパースになる．これが文書クラスタリングが一般のクラスタリングとは異なる特徴であり，単純に一般のクラスタリング手法を文書クラスタリングに用いても良好な結果を得ることは難しい．そのために文書クラスタリングではクラスタリング処理を行う前に主成分分析や特異値分解などの次元縮約の手法を用いることが行われる [3][4]．次元縮約により高次元のベクトルが構造を保った状態で低次元で表現されるため，クラスタリング処理の速度や精度が向上する．

Non-negative Matrix Factorization (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 [15]．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ーの数が 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．

V の各行が X の各列（つまり文書）を k 次元に次元縮約した結果である．ここからクラスタリングを行ってもよいが，NMF では次元縮約した結果自体がクラスタリング結果を表している．V の列の次元は各クラスターのトピックに対応しているからである．つまり基本的な原理は Probabilistic Latent Semantic Indexing (PLSI)[10] と本質的に同じである [9]．また NMF の最大の特徴は，特異値分解を用いる Latent Semantic Indexing (LSI)[4] とは異なり，V の列ベクトルの次元に直交性を必要としていないことである．その結果，NMF では語の出現パターンの類似した文書の集合に対応するように軸が設定され，効果的な文書クラスタリングが行える．また行列 V と U は，ある単純な繰り返し処理から得られるので [12]，LSI のように固有値を求める必要がなく実装も容易という長所がある．

NMF は文書クラスタリングに対して効果的な手法であるが，実際に利用するには以下の２つの問題がある．

- 行列 V と U を得る繰り返し処理の初期値をどのように定めるか．
- 行列 V と U を得る繰り返し処理の終了をどのように判定するか．

実際に，この２点の定め方で NMF によるクラスタリング結果は大きく異なる．

本論文では上記２点の対応策として，min-max cut (Mcut)[8] とその評価関数を利用することを試みる．またここではこの手法を Mcut+NMF と呼ぶことにする．実験では 19 個の文書データセットを用いて，既存手法と Mcut+NMF との比較実験を行い，Mcut+NMF の有効性を示す．

2 NMF とその問題

2.1 NMF とその特徴

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 [15]．$X = UV^T$．ただし k はクラスター数である．

NMF はクラスター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．

NMF の特徴は以下の３点にまとめられる．

- 行列 V と U の要素は非負値である．
- 行列 V はクラスタリング結果を表す．
- V や U に直交行列であるという制約を入れない．

2.2 NMF のアルゴリズム

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 [12]．

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

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

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，繰り返しの回数を決めておくか，UV^T と X との距離 J から判定する．

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

2.3 NMF の問題点

問題 1 行列 V と U の初期値をどのように定めるか.

通常，行列 V と U の初期値にはランダムな値を与える．しかし式 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 と U の初期値の与え方によって，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結果も大きく異なる．

例えば，図 1 は本論文の実験で用いた文書データセット tr45 に対して，NMF によるクラスタリングの実験を 20 回行った結果である．ただし各実験での NMF の初期値にはランダムな値を与えており，各実験の初期値は異なる．図 1 の横軸は実験の番号を示し，縦軸はクラスタリングの精度を表している．図 1 から初期値によって得られる精度が大きく異なることが確認できる．このためできるだけ良好なクラスタリング結果を得るための V と U の初期値を設定する必要がある．

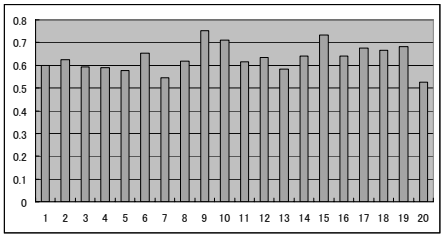

図 1: 初期値とクラスタリングの精度

問題 2 行列 V と U を得る繰り返し処理の終了をどのように判定するか.

通常繰り返し処理の終了の判定のために，式 4 の J の値（あるいは変化量）に閾値を設ける．しかし J の値は分解の近似の良さを表すものであり，直接的にはクラスタリングの精度に対応していない．実際に J の値を小さくしても，クラスタリングの精度が向上せずに，逆に悪化する場合もある．

例えば図 2 は再び文書データセット tr45 に対して NMF を用いた実験結果である．初期値としてはある適当な値を用いている．横軸が NMF の繰り返し回数を表し，LINE-1 が J の値の推移を表し，LINE-2 がクラスタリングの精度の推移を表している．図 2 からわかるように NMF により J の値は減少してゆく，つまり分解の精度は高くなってゆくが，それが直接クラスタリングの精度を高めているわけではないことが確認できる．このためどのような条件で繰り返し処理を終了させるかが，最終的なクラスタリングの精度に大きく関係する．

3 Mcut

本論文は前章で述べた NMF の問題を解決するために Mcut+NMF を提案する．Mcut+NMF で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[8] とその評価関数を利用する．本章では Mcut について解説する．

グラフスペクトル理論を用いたクラスタリングでは，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．この評価関数の最適解がある固有値問題の解に対応することを利用して，クラスタリングを行うのがグラフスペクトル理論を用いたクラスタリングである．評価関数はいくつか提案されているが，ここでは Mcut で提案されているものを利用する．

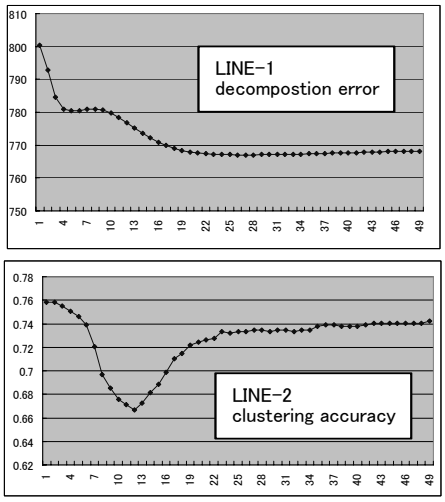

図 2: J の値とクラスタリングの精度

まずサブグラフ 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 間にあるエッジの重みの総和である．エッジの重みはノード（データ）間の類似度を表すので，結局，関数 $W(A,B)$ はサブグラフ A と B の類似度を表している．また，$W(A) = W(A,A)$ と定義しておく．

Mcut の評価関数は以下の式 6 である．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．

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

グラフスペクトル理論を用いたクラスタリングは 2 つのクラスターに分割するのが基本である．目的のクラスター数を得るまで，上記の処理を再帰的に繰り返す．

4 NMF の初期値と終了の判定

NMF では初期値周辺の局所解に収束する，つまり，初期値を多少改善した結果が得られると考えられる．このため初期値としては，かなり良い結果自体を与えるのが効果的だと考えられる．そこで本論文では，最初に Mcut でクラスタリングを行い，その結果から V と U の初期値 V_0 と U_0 を構築する．

具体的には Mcut の結果からデータ i がクラスター c と判定された場合，以下によって V_0 の第 i 行のベクトルを構築する．

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

また U_0 は XV_0 によって構築する．

次に繰り返し処理の終了の判定であるが，これは Mcut により k クラスター $G_1, G_2, \cdots, G_k$ に分割する場合の以下の評価関数を利用する．この評価関数は値が低いほどよい．$\bar{G}_i$ は集合 G_i の補集合を表す．

$$Mcut_K = \frac{cut(G_1, \bar{G}_1)}{W(G_1)} + \frac{cut(G_2, \bar{G}_2)}{W(G_2)} + \cdots + \frac{cut(G_k, \bar{G}_k)}{W(G_k)} \quad (7)$$

具体的には NMF の各繰り返しの終了時にそのクラスタリング結果から式 7 の値を求め，3 回連続してそこまでの最小値が更新されなかった場合に，そこまでの最小値を与えるクラスタリング結果を出力とする．

5 実験

ここでの実験では CLUTO のサイト[1]で公開されている文書データセットを用いる．全部で 24 セットあるが，データ数が 5000 以下である表 1 の 19 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．

表 1: Document Data Sets

Data	文書数	語彙数	クラス数
cacmcisi	4663	41681	2
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
mm	2521	126373	2
re0	1504	2886	13
re1	1657	3758	25
reviews	4069	126373	5
tr11	414	6429	9
tr12	313	5804	8
tr23	204	5832	6
tr31	927	10128	7
tr41	878	7454	10
tr45	690	8261	10
wap	1560	6460	20

まず従来の NMF によりクラスタリングした結果を表 2 の左側に示す．繰り返しを 50 回に固定し，初期値はランダムに与えた．この実験を 20 回行い，最高精度 (Max)，最低精度 (Min)，平均精度 (Mean) および式 4 の評価値から判断して結果を取り出した場合の精度 (NMF-1) を示す．さらに本論文で用いている式 7 の評価値から判断して結果を取り出した場合の精度 (NMF-2) も示す．

各文書データセットに対して，Max と Min の値は大きく異なり，クラスタリングの精度が初期値に依存していることがわかる．また式 4 の評価値から判断して取り出した結果が，必ずしも最高値とはならない．それどころか平均よりも精度が悪く，cacmcisi や k1a は最低値となっている．つまり式 4 の評価値からクラスタリングの精度を見積もるのは有効ではない．一方，本論文で用いた式 7 の評価値を用いた場合，平均よりも精度がよい．これにより，式 7 の評価値でクラスタリングの精度を見積もる方が有効であることがわかる．

次に Mcut+NMF の実験結果を表 2 の右側に示す．また比較として NMF，CLUTO，Mcut の結果も示す．

ここで CLUTO について注記しておく．CLUTO は強力なクラスタリングツールであり，`http://glaros.dtc.umn.edu/gkhome/views/cluto` で公開されている．クラスタリング手法や類似度関数を様々に設定できるが，ここでは default の設定である k-way clustering と呼ばれる手法と cosine の類似度を用いた．多くの実験から経験的に文書クラスタリングに対しては，k-means よりもよい精度がでることが確認されている．

Mcut の精度と比較すると，19 文書データセット中 6 セットは NMF を行ったことで精度が下がったが，7 セットは NMF を行ったことで精度が向上した．残り 6 セットは精度が変化なかった．クラスタリングの平均精度でみると，NMF は 53.50%，CLUTO は 58.21%，Mcut は 61.82% であったのに対し，Mcut+NMF は 63.22% と，最も高い精度を示した．

6 考察

Mcut+NMF で得られるクラスタリング結果の評価関数の値は，Mcut で得られるクラスタリング結果の評価関数の値よりも悪くなることはない．しかし表 2 に示すとおり，Mcut+NMF は Mcut よりもクラスタリングの精度が低くなる場合がある．これは利用した評価関数が，クラスタリングの良さを厳密には表現できていないからである．別の評価関数を用いても，同じ問題は発生する．

クラスタリングのタスクは，評価関数の設定とその関数の最適解の探索方法がポイントである．Mcut+NMF では評価関数としては式 7 を用い，探索方法としては Mcut の探索手法と NMF による探索手法を組み合わせた探索手法として捉えることができる．

近年，グラフスペクトル理論によるクラスタリングを別の枠組みでとらえ直す研究が行われている．例えば Dhillon らはグラフスペクトル理論の評価関数の探索を，weighted kernel を用いることで，k-means の手法で行うことができることを示した [6]．また Ding らはグラフスペクトル理論によるクラスタリングを NMF の枠組みでとらえ直している [7]．このような研究成果を用いれば，統一した枠組みで最適解への探索手法が構築できる．

ただしこのような統一した枠組みによる探索手法では，局所最適解にしか収束しない．そのため局所最適解からよりよい局所最適解へ飛び出す仕組みを入れる方が効果的である．本手法のように異なる探索手法を組み合わせるのも，そのような戦略の 1 つと考えられる．

また Dhillon らが提案した local search [5] は本手法と関連性がある．local search ではまず k-means によってクラスタリングを行い，first variation と呼ぶ手法を用いて，その結果を改善する．そして改善された結果を初期値として再び k-means を実行する．このように k-means と first variation による改善を交互に繰り返す．Mcut+NMF では，まず Mcut によってクラスタリングを行い，NMF によってその結果を改善する．ただし Mcut と NMF の交互の繰り返しはできない．これは Mcut への入力がクラスタリングの結果ではないからである．ただし先に述べた weighted kernel を用いれば，交互の繰り返しも可能となる．

最後に文書クラスタリングの課題について述べたい．

[1] `http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download`

表 2: NMF, CLUTO, Mcut および Mcut+NMF の精度

Data	Max	Min	Mean	NMF-1	NMF-2	CLUTO	Mcut	Mcut + NMF
cacmcisi	0.7667	0.5788	0.6265	0.5788	0.6030	0.6054	**0.6858**	**0.6858**
cranmed	0.8420	0.4648	0.6601	0.5825	0.7133	**0.9975**	0.9930	0.9930
fbis	0.5112	0.3707	0.4374	0.4125	0.4296	0.4921	**0.5278**	0.4941
hitech	0.4689	0.3364	0.4066	0.4633	0.4124	**0.5228**	0.3859	0.5059
k1a	0.5274	0.4107	0.4673	0.4107	0.5197	0.4799	0.4658	**0.5684**
k1b	0.8530	0.5457	0.6969	0.6389	**0.7222**	0.6081	0.5205	0.5342
la1	0.6807	0.5069	0.6060	0.6798	0.6807	**0.7147**	0.6879	0.6879
la2	0.7232	0.4198	0.5717	0.5873	0.5857	0.6582	**0.7028**	0.6924
mm	0.6418	0.5200	0.5640	0.5470	0.5303	0.5331	**0.9583**	0.9556
re0	0.4648	0.3358	0.4021	**0.3710**	0.3358	0.3198	0.3670	0.3670
re1	0.4267	0.3458	0.3904	0.3826	0.4049	0.4146	0.4490	**0.4599**
reviews	0.7569	0.4726	0.5815	**0.7196**	0.5353	0.6316	0.6776	0.6424
tr11	0.6763	0.4758	0.5850	0.5556	0.5797	0.6812	0.6546	**0.7295**
tr12	0.6645	0.4728	0.5867	0.6422	0.6422	0.6869	**0.7764**	**0.7764**
tr23	0.5294	0.2941	0.4333	0.3971	**0.5294**	0.4559	0.4363	0.4363
tr31	0.6311	0.4196	0.5196	0.5696	0.5275	0.5674	**0.7228**	0.6624
tr41	0.6503	0.4875	0.5936	0.5239	0.6059	**0.6412**	0.5661	0.6014
tr45	0.7507	0.5261	0.6327	0.6347	0.6754	0.5986	**0.7580**	0.7101
wap	0.5250	0.3904	0.4621	0.4686	0.4654	0.4487	0.4109	**0.5096**
平均	0.6363	0.4408	0.5381	0.5350	0.5525	0.5821	0.6182	**0.6322**

クラスタリングの問題はデータをいったんベクトル表現した後は，純粋に工学的な問題となる．しかしより精度の高いクラスタリング結果を得るためには，ベクトル表現する前の段階の知識をより積極的に利用すべきである．

クラスタリングは教師なし学習である．精度を求めるのであれば，教師を付けた方がよい．近年ユーザーとのインタラクションを用いた semi-supervised なクラスタリングが活発に研究されている [2][1][14]．メタな情報を教師として利用し，semi-supervised な枠組みでクラスタリングする方向が有望である．

7 おわりに

本論文では NMF の繰り返し処理における初期値と終了の判定の問題を提示し，その対策として Mcut の結果から初期値を構成し，Mcut の評価関数を終了判定に用いる Mcut+NMF を提案した．実験では，NMF と Mcut 及び CLUTO との比較実験を行い，Mcut+NMF の有効性を示した．文書クラスタリングに特化したメタ知識の導入と semi-supervised のクラスタリングの実現を今後の課題とする．

参考文献

[1] Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. Semi-supervised Clustering by Seeding. In *ICML-2002*, pp. 19–26, 2002.

[2] Mikhail Bilenko, Sugato Basu, and Raymond J. Mooney. Integrating Constraints and Metric Learning in Semi-Supervised Clustering. In *ICML-2004*, pp. 81–88, 2004.

[3] Daniel Boley, Maria L. Gini, Robert Gross, Eui-Hong Han, Kyle Hastings, George Karypis, Vipin Kumar, Bamshad Mobasher, and Jerome Moore. Document categorization and query generation on the world wide web using webace. *Artificial Intelligence Review*, Vol. 13, No. 5-6, pp. 365–391, 1999.

[4] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, Vol. 41, No. 6, pp. 391–407, 1990.

[5] 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.

[6] Inderjit Dhillon, Yuqiang Guan, and Brian Kulis. A Unified View of Kernel k-means, Spectral Clustering and Graph Cuts. In *The University of Texas at Austin, Department of Computer Sciences. Technical Report TR-04-25*, 2005.

[7] Chris Ding, Xiaofeng He, and Horst D. Simon. On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. In *SDM 2005*, 2005.

[8] 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.

[9] 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.

[10] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pp. 50–57, 1999.

[11] Krishna Kummamuru, Rohit Lotlikar, Shourya Roy, Karan Singal, and Raghu Krishnapuram. A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results. In *Proceedings of WWW-04*, pp. 658–665, 2004.

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

[13] Michael W. Berry, editor. *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer, 2003.

[14] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, pp. 505–512, 2003.

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