動画共有サイトにおけるユーザ投稿コメント解析

澤田　敬治† 　手塚　太郎†† 　木村　文則†† 　前田　亮††

† 立命館大学 理工学研究科　〒 525–8577 滋賀県草津市野路東 1–1–1
†† 立命館大学 情報理工学部　〒 525–8577 滋賀県草津市野路東 1–1–1
E-mail: †cm005053@ed.ritsumei.ac.jp, ††{tezuka,amaeda}@media.ritsumei.ac.jp,
†††fkimura@is.ritsumei.ac.jp

あらまし　近年，多数の動画投稿サイトが存在し，膨大な数の動画がユーザに提供されており，動画に対してコメントを残せるサイトも存在する．これらのコメントは，映像を見ることなく動画の内容を知るための手掛かりとなるものである．そこで本研究では，動画に対して投稿されたコメントをクラスタリングする手法について述べる．クラスタリングには，高次元の行列を低次元の行列に変換することのできる，非負値行列因子分解 (Non-negative Matrix Factorization：NMF) と確率的潜在意味解析 (Probabilistic Latent Semantic Indexing：PLSI) とのハイブリッド法によってクラスタリングする手法を採用する．この手法により NMF のみのクラスタリングよりも精度を向上させることができた．

キーワード　クラスタリング，NMF，PLSI，ハイブリッド，コメント

User Contribution Comments Analysis of Video Sharing Site

Keiji SAWADA†, Taro TEZUKA††, Fuminori KIMURA††, and Akira MAEDA††

† Graduate School of Science and Engineering, Ritsumeikan University　1-1-1 Noji-higashi，Kusatsu, Shiga 525–8577, Japan
†† College of Information Science and Engineering　1-1-1 Noji-higashi，Kusatsu, Shiga 525–8577, Japan
E-mail: †cm005053@ed.ritsumei.ac.jp, ††{tezuka,amaeda}@media.ritsumei.ac.jp,
†††fkimura@is.ritsumei.ac.jp

Abstract　A lot of video sharing sites exist, a great number of video contents are offered to the user, and the sites where the comment can be submitted to video contents exists. These comments become clues to know of the content of video without seeing the video itself. In this study, the method for the technique for clustering is described about the comment contributed to video. We used the hybrid method of NMF (Non-negative Matrix Factorization) and PLSI(Probabilistic Latent Semantic Indexing) for clustering. The method converts a higher dimension matrix into a lower one and finds topics in comments. Accuracy improved by this technique compared to clustering using NMF only.

Key words　Clustering, NMF, PLSI, Hybrid, Comments

1. はじめに

近年，インターネット上での動画共有が盛んになってきており，特に動画上に直接コメントを書き込めるシステムが多くの注目を集めている．これらのシステムでは動画という画像による情報だけでなく，コメントという言語による情報も得ることができる．言語情報は表現力が豊かであるため，画素や輝度等を使った画像情報による分類よりもより細かな分類が可能になるのではないかと考える．

そこで本研究では，この言語情報に着目しコメントをクラスタリングすることで動画の分類を行う．動画に付けられたコメントに対して NMF と PLSI とのハイブリッド法を適用することで，コメントのクラスタリングを行う．この手法によって得られるクラスタリングでは，動画の内容だけではなく閲覧者目線も考慮した分類が行える可能性がメリットとして考えられる．

動画上に直接コメントを書き込めるシステムの一つとして現在広く利用されているのは，ニコニコ動画 [1] であるので本研究ではニコニコ動画を対象として実験を行う．ニコニコ動画には約 370 万の動画と約 23 億ものコメントが蓄積されている (2010 年 2 月現在) ので，研究対象としては十分な情報量が得

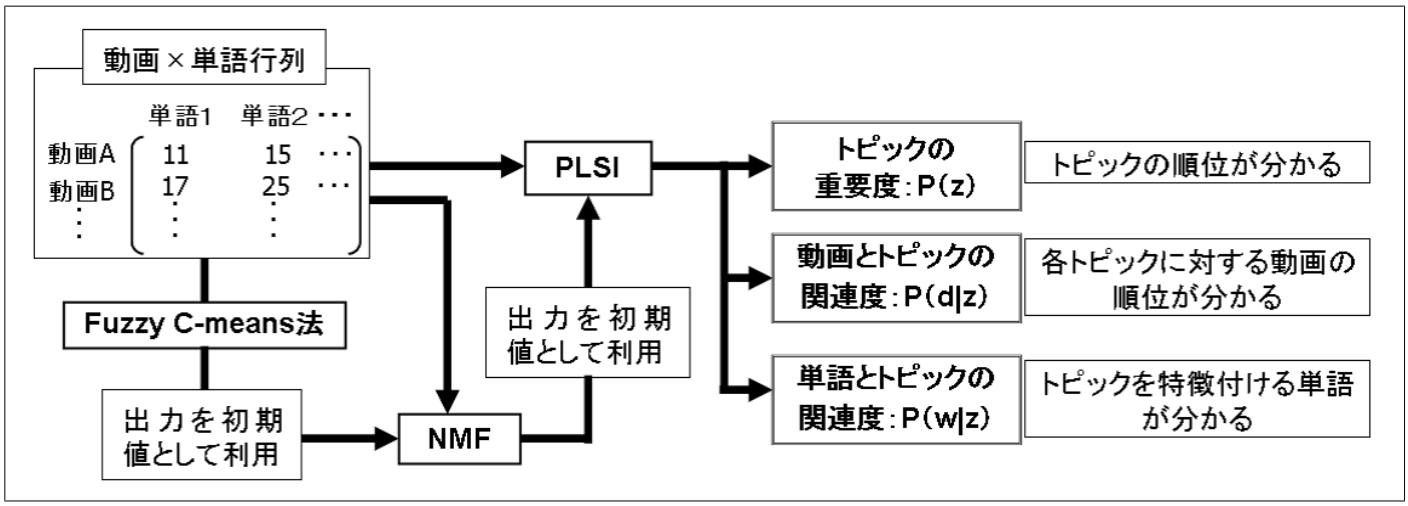

図 1　ハイブリッド法の流れ：「動画×単語行列」に Fuzzy C-means 法を実行した結果を初期値として「動画×単語行列」に対して NMF を行う．その結果を初期値として「動画×単語行列」に対して PLSI を行うことで「トピックの重要度 ($P(z)$)」「動画とトピックの関連度 ($P(d|z)$)」「単語とトピックの関連度 ($P(w|z)$)」を得る．

られると考える．

2. 関 連 研 究

文書分類の手法にはナイーブベイズ分類器や SVM(Support Vector Machine) などの教師あり学習と，NMF や PLSI などの教師なし学習が存在する．教師あり学習では，分類先が決まっているため分類誤差が少ないという利点があるが，学習していない未知の入力データに対しては弱いという側面を持っている．一方，教師なし学習では事前に分類先を決めていないので，未知の入力データに対しても対応することができる．本研究におけるデータは，個々人がその場での感情を表したのもであるので入力データが不規則になったり，誤字が含まれる可能性が考えられるので，教師なし学習を用いる．

2.1 NMF

NMF(Non-negative Matrix Factorization：非負値行列因子分解) は，高次元の行列を低次元の行列に圧縮するための手法である．具体的には非負の行列 $Y(n \times m)$ を，非負の行列 $W(n \times r)$ と $H(r \times m)$ の 2 つの行列に分解するというものである $(r < n, m)$(式 1)．

$$Y \approx WH \tag{1}$$

これにより，n 行あった行列を r 種類の特徴で表すことができるようになる [2]．また，このとき得られた 2 つの行列 W，H はそれぞれ，動画の特徴を表す単語の行列 (特徴行列) とその特徴を持つ動画の寄与率を表す行列 (寄与率行列) となる．

NMF ではランダムに選ばれた W，H から出発し，反復法によって Y と WH の相違を最小化する．相違を測る尺度としてユークリッド距離とカルバック・ライブラー (KL) 情報量が提案されており [3] [4]，本研究では後者を採用している．その方法は式 2，式 3 の通りであり，

$$\overline{H}_{ij} = H_{ij} \sum_{k}^{r} W_{ki} \frac{Y_{kj}}{(WH)_{kj}} \tag{2}$$

$$\begin{aligned} \hat{W}_{ij} &= W_{ij} \sum_{k}^{r} \frac{Y_{ik}}{(WH)_{ik}} H_{jk} \\ \overline{W}_{ij} &= \frac{\hat{W}_{ij}}{\sum_{k}^{r} \hat{W}_{kj}} \end{aligned} \tag{3}$$

この計算を繰り返し行い W と H を更新する．ここで，$\overline{H}$，$\overline{W}$ はそれぞれ更新された H と W を示す．

以上の更新式は Y と WH の間の KL 情報量を元にした以下の目的関数 (式 4) を単調減少させることが証明されている [2]．

$$E = \sum_{i}^{m} \sum_{j}^{n} (Y_{ij} \log \frac{Y_{ij}}{(WH)_{ij}} - Y_{ij} + (WH)_{ij}) \tag{4}$$

2.2 PLSI

PLSI(Probabilistic Latent Semantic Indexing：確率的潜在意味解析) は，確率モデルに基づき潜在的な意味の関連性を見つけ出す手法である [5] [6]．つまり，ある文書 d とある単語 w との間にある関係，すなわち 2 つの共通のトピックとなるような特徴 z を見つけ出すことが PLSI の目的である．そして，文書 d，単語 w，トピック z の同時確率は以下式 5 のように周辺化できる．

$$P(w, d) = \sum_{k}^{K} P(w, d, z_k) \tag{5}$$

また，条件付き確率の定義より式 7 が，w と d が z のもとで条件付き独立であるという仮定を設けることで式 7 が言える．

$$P(w, d, z) = P(w, d \mid z) P(z) \tag{6}$$

$$P(w, d \mid z) = P(w \mid z) P(d \mid z) \tag{7}$$

これらの式をまとめると以下が得られる．

$$P(w, d) = \sum_{k}^{K} P(w \mid z_k) P(d \mid z_k) P(z_k) \tag{8}$$

式 8 において各分布のパラメータは，文書 d に含まれる単語 w の数 $n(d,w)$ を観測変数として用いたで最尤法で求められる．尤度の最大化と同義の対数尤度 L の最大化を行う．

$$L = \sum_{i}^{N}\sum_{j}^{M} n(d_i, w_j) \log \sum_{k}^{K} P(w \mid z_k)P(d \mid z_k)P(z_k) \quad (9)$$

式 9 を目的関数とし，EM アルゴリズムにより対数尤度 L の最大化を行う．

2.3 ハイブリッド法

NMF と PLSI は更新式を用い，解へと収束するという点では同じであるが，これらは全く別のアルゴリズムである．しかし，Ding らによって，I-divergence(Information divergence, 相対エントロピー，KL 情報量) による NMF と PLSI は別のアルゴリズムにもかかわらず，同じ目的関数を最適化していることが示された [7]．更に，NMF と PLSI は共に高次空間において局所解に陥る可能性があるのだが，NMF が陥った局所解から PLSI により抜け出すことができ，またその逆も可能であることも示されている．このことから，NMF と PLSI を交互に実行するハイブリッド法によってそれぞれが陥る局所解から抜け出し，より良い解へと近づくことができるので，それぞれの手法単独での処理よりも精度を向上させることができる．

3. 提 案 手 法

本研究では，ニコニコ動画から API(Application Programming Interface) を用いてコメントを取得し，そのコメントを形態素解析したものを使用する．次に，形態素解析によって得られた単語を動画ごとに並べた「動画×単語行列」を作り，その行列に対してクラスタリング (ハイブリッド法) を行う．ハイブリッド法を行う際には，最初に k-means 法を実行する．これは，NMF・PLSI の両手法が共に初期値依存性を有しているためで，k-means 法を用いて初期値を決定することでその問題を解消できるからである．本研究では，NMF と PLSI が EM アルゴリズムを使用したソフトクラスタリングであるため，ハードクラスタリングである k-means 法の代わりに Fuzzy C-means 法を用いている．また，Fuzzy C-means 法を採用することによりクラスタリング結果のスパース性を軽減することができるので，全ての要素に小さな定数を足すというスムージングの操作をしなくてよいという利点もある．この Fuzzy C-means 法のクラスタリング結果を初期値として NMF を実行する．そして，NMF のクラスタリング結果を初期値として PLSI を実行し，最終的な結果を得る．

ハイブリッド法の流れを図 1 に示す．PLSI によるクラスタリングの結果，$P(d|z)$, $P(w|z)$, $P(z)$ が得られる．$P(w|z)$ はどのような単語群がトピックとなるかを，$P(z)$ はどのトピックが重要かを表している．そして，$P(d|z)$ はどのトピックからどの文書 (動画) が得られるかを表している．

よって，まず $P(z)$ から重要なトピックを見つけ，そのトピックの素となる単語群を $P(w|z)$ により取り出す．最後にトピックに対応する $P(d|z)$ の文書 (動画) が，$P(w|z)$ の単語群とマッチしているかを比較することで評価を行う．

表 1　NMF のみによる実験の評価結果および適合率

正解	U1	U2	U3	U4	U5
a1	x	a	x	x	d
a2	d	d	a	a	a
a3	-	-	a	x	-
b1	a	d	b	b	x
b2	a	a	b	b	d
b3	-	-	b	b	-
c1	x	c	b	b	d
c2	d	d	a	x	x
c3	c	c	c	x	x
d1	d	d	d	x	x
d2	x	a	d	d	x
d3	x	d	d	x	x
適合率	0.167	0.417	0.750	0.417	0.083

表 2　ハイブリッド法による実験の評価結果および適合率

正解	U1	U2	U3	U4	U5
a1	x	a	x	a	a
a2	a	a	a	a	a
a3	-	-	a	x	-
b1	b	d	b	x	b
b2	x	a	a	x	x
b3	x	b	b	x	x
c1	x	b	c	x	x
c2	c	c	c	x	c
c3	c	c	c	b	c
d1	-	-	d	a	-
d2	d	d	d	d	d
d3	d	d	d	x	-
適合率	0.500	0.583	0.833	0.250	0.500

4. 実験と評価

評価は NMF とハイブリッド法それぞれで行い，両手法での精度を比較する．評価方法は，NMF ではトピックとなる単語群を表す $P(z)$ に対応する出力を得られないので，トピックとなる単語群は人手により 4 つ取り出し，ハイブリッド法では重要なトピック順に 4 つの単語群を取り出す．そして，各トピックに対応する動画を 3 つずつ計 24 動画取り出す．取り出したトピックをそれぞれ「a〜d」と名付け，各トピックの動画を「a1, a2, a3, …」とする．取り出した動画を 5 人の被験者が閲覧し，その動画がどのトピックとなる単語群に近いかを評価してもらう．表 1 が NMF のみ，表 2 がハイブリッド法の評価結果であり，表中の U はユーザを表している．また，ユーザ列の「a〜d」はユーザが各動画がどのトピックに最も相応しいかを評価した結果を表し,「x」はどのトピックとも当てはまらない,「-」は動画が削除されていたため評価できなかったことを表している．評価の良し悪しは適合率によって表し，正解とユーザの評価が同じものの数を，視聴した動画数で割ることで求める．

5. 考　　察

表 1 および表 2 を見ると，5 ユーザ中 4 ユーザの評価におい

て適合率が上昇していることが分かる．この結果から，ハイブリッド法によるクラスタリングは，NMF のみによるクラスタリングよりも適合率を向上させることが確認できた．適合率が向上したユーザに限って言えば，ハイブリッド法での適合率が 5 割を超えることができた．

6. ま と め

本研究では，NMF と PLSI によるハイブリッド法を用いて，動画に付けられたコメントに対してクラスタリングを行うことで動画をクラスタリングする手法を提案した．この結果，ハイブリッド法は日本語のコメントの分類にも有用であることが確認され，NMF のみのクラスタリングよりも良い適合率を得た．また，NMF のみのクラスタリングでは，どのクラスタに割り当てるか判断に困る物 (動画) が多かったが，ハイブリッド法でのクラスタリングでは，即座に振り分けることができる物が多くなった．このことから，ハイブリッド法ではトピックを特徴付ける単語群がより適切に選ばれていると考えることができる．

将来的には，品詞の使用条件や投稿時間等の他のファクタをクラスタリングへと反映させることができれば，これまでの動画分類とは違った切り口による動画の分類が可能になるのではないかと考える．

謝　　辞

本研究の一部は文部科学省私立大学戦略的研究基盤形成支援事業「芸術・文化分野の資料デジタル化と活用を軸とした研究資源共有化研究」の支援を受けている．

文　　献

[1] ニコニコ動画，http://www.nicovideo.jp/
[2] e Daniel D. Lee，H. Sebastian Seung，“Algorithms for Non-negative Matrix Factorization”，Advances in Neural Information Processing Systems Vol.13，pp.556-562 (2001).
[3] e Chih-Jen Lin，“Projected Gradient Methods for Non-negative Matrix Factorization”，Neural Computation，Vol.19，No.10，pp.2756-2779 (2007).
[4] 柘植 覚，獅々堀 正幹，北 研二，Non-negative Matrix Factorization を用いた情報検索，IPSJ SIG Notes Vol.2001，No.20 pp.1-6，(2001).
[5] e Thomas Hofmann，“Probabilistic latent Semantic Analysis”，Proc. of the 15th Conference on Uncertainty in Artificial Intelligence，UAI-99．Morgan Kaufmann Publishiers，San Francisco，CA，pp.289-296(1999)
[6] e Thomas Hofmann，“Probabilistic Latent Semantic Indexing”，Proc. of SIGIR ’99，ACM Press，pp.50-57(1999)
[7] e Chris Ding，Tao Li，Wei Peng，“On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing”，Computational Statistics & Data Analysis Volume 52，Issue 8，pp.3913-3927(2008)