マルチメディアデータに対する周辺情報を用いた Latent Dirichlet Allocation によるタグ付け支援

津田 覚之 † 三輪 誠 ‡ 鶴岡 慶雅 † 近山 隆 †

† 東京大学 工学部 電子情報工学科 ‡ the University of Manchester, UK

†{tsuda, tsuruoka, chikayama}@logos.t.u-tokyo.ac.jp
‡makoto.miwa@manchester.ac.uk

1 はじめに

近年，ユーザが動画・画像・音楽などのマルチメディアデータを投稿し，他ユーザに公開することができるような Web サービスが広く普及している．そのうちニコニコ動画[1]などの多くのサイトで，任意のユーザが付けることができるタグによる分類がなされており，タグを用いた検索が可能となっている．

一方，タグ付けに関する個々のユーザの判断基準はそれぞれ異なるため，付けられることが望ましいようなタグが付いていなかったり，タグの表記揺れが発生したりするという問題がある．

本研究では，マルチメディアデータの投稿物に対して，適切と思われるタグの候補を提示することを目的とする．その際，マルチメディアデータの内容を読み取ることは難しく計算量が大きいということに加え，タグの付けられていない投稿物に対応するため，タイトルや投稿者による説明文など，文字の形で付随する周辺情報を用いる．

[4] や [5] といった先行研究では，既に付けられているタグに基づいてタグ推薦が行われた．本研究では投稿物の周辺情報に対して [5] と同様に LDA を用い，その属するトピックごとの重みを得ることで，より良いタグ推薦を行うことを目指す．

2 関連研究

2.1 Latent Dirichlet Allocation (LDA)

Blei らによる Latent Dirichlet Allocation (LDA, 潜在的ディリクレ配分法)[2] は，文書集合からトピック（話題）を発見するための統計的なモデル（トピックモデル）の 1 つである．各文書が潜在変数を持つと考え，文書の内容を最もよく特徴付けるような潜在変数を求める．

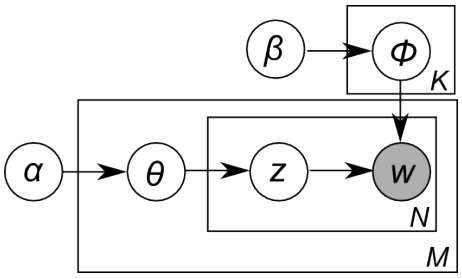

図 1: LDA

以下，文書集合 M, トピック集合 K, 語彙集合 V とする．LDA は図 1 のように図示され，以下のアルゴリズムで示される．

1. ハイパーパラメータ α のディリクレ分布によって $M \times K$ 行列 θ を生成する．
2. ハイパーパラメータ β のディリクレ分布によって $K \times V$ 行列 ϕ を生成する．
3. 各文書 $i \in M$ の中で j 番目の位置にある各単語について，
   (a) θ_i の多項分布によってトピック $z_{ij} \in K$ を生成する．
   (b) $\phi_{z_{ij}}$ の多項分布によって単語 $w_{ij} \in V$ を生成する．

各文書の各単語についてこれを繰り返し，Collapsed Gibbs Sampling や Collapsed Variational Bayes などの手法によって各パラメータの推論を行う．

[1] http://www.nicovideo.jp/video_top/

2.2 LDA を用いたタグ推薦

Krestel らは，ソーシャルブックマークサービス Delicious[2] のデータセットでの，既に付けられているタグに基づくタグ推薦において，LDA を用いたタグ推薦手法が，組み合わせ規則を用いた Heymann らのタグ推薦手法 [4] より良い結果を出すことを示した [5].

3 提案手法

マルチメディアデータの投稿物に対して，[4] や [5] のように既に付けられているタグを用いてではなく，タイトルや説明文などの周辺情報を用いてタグの推薦を行うことを目的とする．周辺情報はタグと比較して，トピックを 1 つに決めることが適当でない単語も多いと考えられるため，LDA を用いてトピックごとに重み付けを与えることでタグ推薦を行う．それによって，動画の周辺情報が持つトピック分布の情報に基づいた，適切なタグが推薦されることを期待する．

以下の手順によってタグ推薦を行い，手法 1〜3 について結果を比較する．2. における n'_w の各手法での例を表 1 に示す.

1. 文書集合に対して LDA を適用し，各単語が各トピックに割り振られた回数を示す $V \times K$ 行列 n を得る.
2. $V \times K$ 行列 n' の w 行目 (n'_w) を以下のいずれかの手法で定める.
   - **手法 1:** ***Continuous*** n_w を正規化した（和が 1 になるようにした）ベクトル.
   - **手法 2:** ***Discrete*** n_w の最大値を取るトピックで 1，他で 0 となるベクトル.
   - **手法 3:** ***CutOut*** *Continuous* のベクトルの各要素について，ある閾値を下回る値のみ 0 としたベクトル.
3. タグ推薦対象の文書中の各単語 w_{ij} についての $n'_{w_{ij}}$ を加算し，トピックごとの重みベクトルとする.
4. n のうち，タグ以外の単語に対応する行の数値を全て 0 にする.
5. n と得られた重みベクトルとの行列積を得て，その結果の高い順に語を出力する．4. の操作により，タグのみが推薦される.

[2] https://delicious.com/

4 実験と考察

4.1 データセット

ニコニコ動画のデータセット[3]を用いた．動画のタイトル，投稿者による動画説明文，タグなどのデータを用いることができる.

計算量上の理由により，データセット約 830 万件のうち動画の投稿順の先頭から 100 万件を用いた．各動画におけるタイトル・動画説明文・タグの 3 種類のデータをひとまとめにしたものを，LDA における文書として扱った.

その際，タイトルと動画説明文は形態素解析して単語に分割し，不要と思われる語やタグ，出現数 10 以下の語やタグは除いた．データセットに特有の語が正しく分割されやすくするために，形態素解析のための追加辞書としてタグを使用した．また，タグには接頭語を付加し，タイトルや動画説明文に現れる語とは別のものとして扱われるようにした.

4.2 LDA の適用

4.1 のデータセットに対して LDA を適用した．Collapsed Gibbs Sampling[3] によって推論を行う Java 実装を用い，トピック数 $K = 100, 150, 250, 400$，イテレーション回数は 250 回とした．ハイパーパラメータは $\alpha, \beta = 0.5$ を初期値として，fixed-point iteration[1, 6] によって随時更新を行った.

イテレーションに伴うパープレキシティの推移は図 2 のようになった．以降では，パープレキシティの最も低くなった $K = 400$ の結果を使用した.

4.3 LDA を用いたタグ推薦

4.2 の結果を用いて，3 章の手法によってタグ推薦を行った．乱数シードを固定してデータセット 100 万件から 1 万件をランダムに選択し，そのうちタイトルと動画説明文だけをデータとして与えて各 50 個までタグ推薦を行った.

表 1: トピック数 $K = 5$ のときの各手法での n'_w の例

Continuous	0.3	0.4	0.2	0.1	0
Discrete	0	1	0	0	0
CutOut（閾値 0.3）	0.3	0.4	0	0	0

[3] http://www.nii.ac.jp/cscenter/idr/nico/nico.html

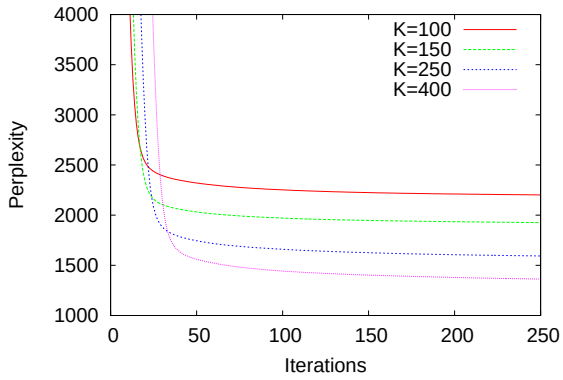

図 2: 各トピック数でのパープレキシティの推移

4.3.1 *Continuous* と *Discrete* との比較

まず，3 章の *Continuous* と *Discrete* を比較した．それぞれの適合率, 再現率, F 値は図 3〜6 のようになった．なお，本研究においての適合率, 再現率は「推薦したタグが対象の動画に実際に付いていたかどうか」を判断基準としている．

大きな差はないが，どのタグ推薦数においても *Continuous* が *Discrete* を上回った．

4.3.2 *CutOut* の閾値による結果の比較

Discrete は本研究の対象に対して 4.3.1 のように欠点が考えられるが，*Continuous* も，多くのトピックに分散する単語に対してそれら全てのトピックに重みを与えるといった動作が不適切な方向に働く可能性が考えられる．

このため，3 章の *CutOut* について，閾値を 0 から 0.9 まで 0.1 ずつ変化させて比較した．それぞれの適合率, 再現率, F 値は図 7〜9 のようになった．なお，閾値 0 では *Continuous* と同じものである．

適合率, 再現率のいずれも，閾値 0.1〜0.3 程度で最大値を取った．閾値を設けない場合と比較して，単語との関連性の薄いトピックに重みを与えないことで性能向上に効果があると考えられる．

5 おわりに

本研究では，マルチメディアデータの投稿物に文字の形で付随する周辺情報を用いて，適切と思われるタグの候補を提示する手法を提案した．

既に付けられているタグに基づいてタグ推薦を行った先行研究と異なり，本研究においては投稿物のタイ

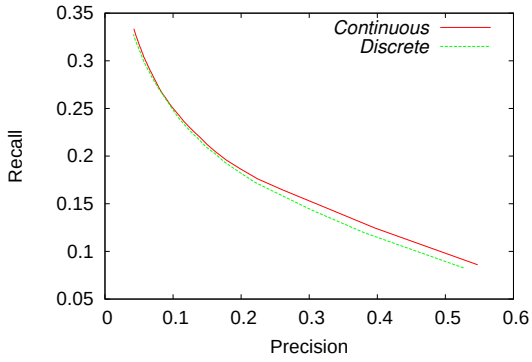

図 3: *Continuous* と *Discrete* の適合率, 再現率

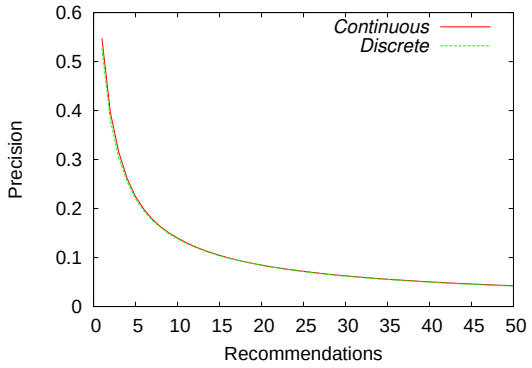

図 4: 両手法のタグ推薦数ごとの適合率

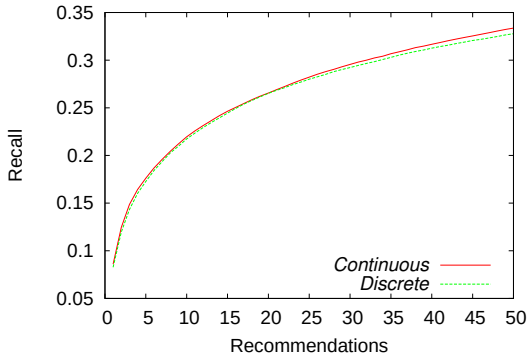

図 5: 同再現率

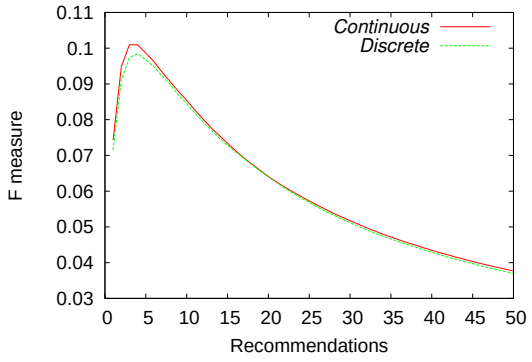

図 6: 同 F 値

トルと説明文という周辺情報に基づいてタグ推薦を行った．周辺情報はタグと比較して，トピックを 1 つに決めることが適当でない単語も多いと考えられるため，LDA を用いてトピックごとに重み付けを与えることでタグ推薦を行った．その際，重み付けに関して複数の手法を比較検討し，単語ごとにトピックに与える重みの閾値を適当に設定することで性能が向上することを示した．

今後の課題としては以下が挙げられる．動画の周辺情報内の単語を（一部を除去したほかは）特に区別せず用いたが，あらかじめ重要語抽出を行えば関連するタグをより効果的に推薦できると思われる．また，視聴者の感想など，周辺情報の持つ内容に直接関係しないタグが存在するため，それを除くことで不正確な推薦を減らせる可能性がある．加えて，適合率と再現率の判定は推薦したタグが付いていたかどうかで行ったが，実際に付いてはいなくても適切と思われるタグも少なくないため，精確な評価は人が実際に見て行う必要があると考えられる．

参考文献

[1] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 27–34. AUAI Press, 2009.

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

[3] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, Vol. 101, No. Suppl 1, pp. 5228–5235, 2004.

[4] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social tag prediction. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 531–538. ACM, 2008.

[5] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pp. 61–68. ACM, 2009.

[6] Thomas Minka. Estimating a dirichlet distribution, 2000.

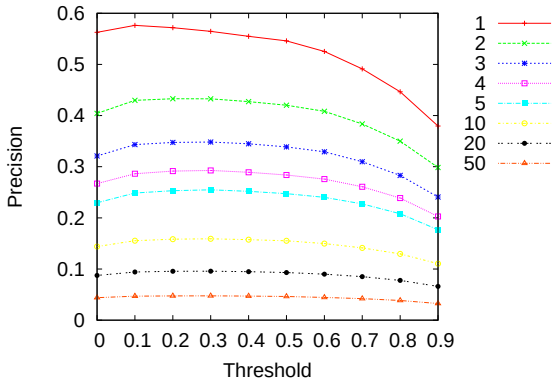

図 7: *CutOut* のタグ推薦数 1〜50 での閾値ごとの適合率

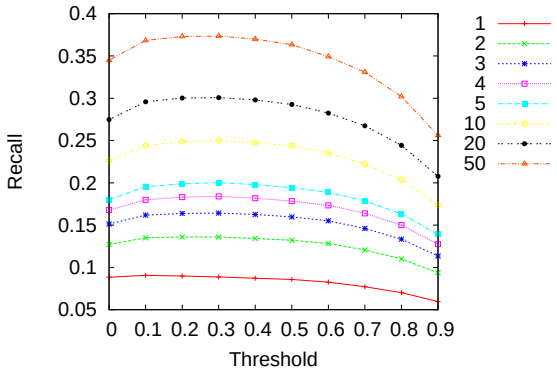

図 8: 同再現率

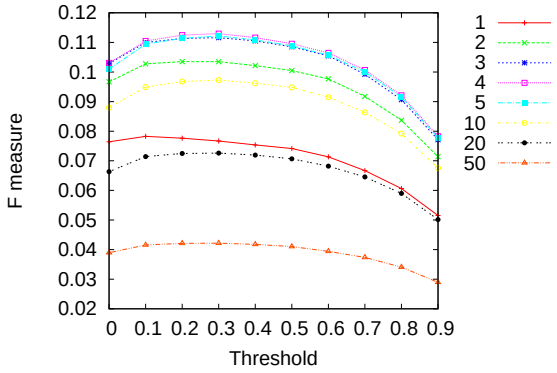

図 9: 同 F 値