時間情報を用いた文書への自動タグ付与モデルに関する検討

加藤 亮　　吉川 大弘　　古橋 武
名古屋大学大学院工学研究科
{katou}@cmplx.cse.nagoya-u.ac.jp
{yoshikawa,furuhashi}@cse.nagoya-u.ac.jp

1 はじめに

近年，インターネットの普及により，膨大な数の文書が発信されており，有益な情報を絞り込むことは容易ではない．そのため，所望する文書を効率的に探し出すための，文書の内容に基づいた知識整理を行う手法が必要であると考えられる．知識整理を行う方法として，整理対象を代表するような短い言葉（タグ）を付与する方法が用いられることが多い．近年では，このタグを文書に自動で付与し，知識整理を行う手法が数多く報告されている [1][2][3][4]．それらの中でも，トピックモデルを用いた研究が近年注目され，また成果を挙げている [5]．

上述のトピックモデルでは，単語とタグの背景にそれぞれトピックを仮定し，確率分布により単語とタグが生成されるとした確率モデルである．文献 [5] では，ブログデータの一部を学習に用い，学習に用いなかったデータをテストデータとしてタグ付与実験を行うことで，従来用いられている手法よりも高い適合率と再現率でタグ付与が行えることを示している．しかし，これらの手法では，単語やタグに対する時間変化を考慮しておらず，時間とともに変化する話題のタグを付与することは困難となると考えられる．

本稿では，文書に対する自動タグ付与を目的として，文書に付与されるタグと，文書に含まれているトピック，文書が生成された時間の 3 つに着目した新しいトピックモデルの構築を目指す．

2 提案モデル

2.1 概要

本稿では Tag-LDA [5] と時系列トピックモデルの一つである Topics over Time (TOT)[6] を統合したモデルを提案する．TOT は時系列文書をモデル化した手法で，連続な時間分布 (timestamps）に対してトピックを仮定したものである．この TOT の手法に対して，時間とともに変化する話題を捉えた，タグ付与のためのトピックモデルを提案する．提案モデルは，単語，タグ，timestamps のそれぞれにトピックを仮定し，またそれらを生成するモデルである．トピックが生成しやすい timestamps を知ることで，話題の盛衰を捉えたタグ付与が可能となることが期待される．

単語とタグのトピックを仮定した確率モデル，あるいは，timestamps を用いた時系列トピックモデルは提案されているものの，それらを組み合わせたトピックモデルについては，これまであまり報告されていない．

2.2 モデル化

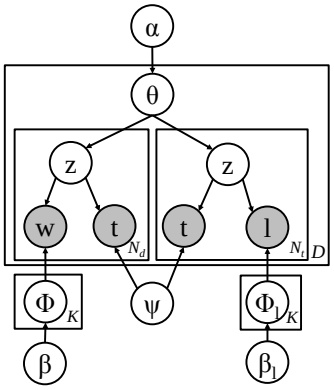

図 1: 提案モデルのグラフィカルモデル

本稿では，上述のトピックモデルをモデル化することを目指す．トピックからタグと単語，timestamps の出現にそれぞれ多項分布を仮定する．提案モデルのグラフィカルモデルを図 1 に示す．図 1 において，D は文書数，K はトピック数，N_t はタグの総種類数，N_d は文書 d 内の単語数を表す．提案モデルでは，以下の仮定により文書集合が生成されるとする．

1. For each topic z

(a) Draw a word distribution $\phi \sim Dir(\beta)$

(b) Draw a tag distribution $\phi_l \sim Dir(\beta_l)$

2. For each document d

(a) Draw a topic distribution $\theta^d \sim Dir(\alpha)$

(b) For each word i in document d

i. Draw a topic $z_i \sim Mult(\theta^d)$

ii. draw a word $w_i \sim Mult(\phi^z)$

iii. draw a timestamps $t_i \sim Beta(\psi^z)$

(c) For each tag i in document d

i. Draw a topic $z_i \sim Mult(\theta^d)$

ii. draw a tag $l_i \sim Mult(\phi^z)$

iii. draw a timestamps $t_i \sim Beta(\psi^z)$

ここで，$Dir(\cdot)$ はディリクレ分布，$Mult(\cdot)$ は多項分布，$Beta(\cdot)$ はベータ分布を表し，α，β，β_l はそれぞれのディリクレ分布におけるハイパーパラメータである．

2.3 トピック・タグの推定

提案モデルでは，LDA や TOT と同様に，ギブスサンプリング [7] を用いてトピックとタグの推定を行う．

ギブスサンプリングでは，位置 i のトピック z_i を位置 t_i 以外の情報を用いて推定する．ギブスサンプリングにおける更新式は以下のように表せる．

$$p(z_i|z_{\setminus i}, t_i, \mathbf{w}, \mathbf{l}) \tag{1}$$

$$\propto \frac{n^l_{v,\setminus i} + \beta}{n^l_{-,\setminus i} + V\beta} \cdot \frac{n^d_{k,\setminus i} + \alpha}{n^k_{-,\setminus i} + K\alpha} \cdot \frac{(1-t_i)^{\psi_{1,k}} \cdot t_i^{\psi_{2,k}}}{B(\psi_{1,k}, \psi_{2,k})} \tag{2}$$

ただし，上式における “$\setminus i$” は，i 番目の要素を除いた場合のカウントを示す．式 (2) に基づくサンプリングを十分回数実行することによって，潜在的な変数が推定される．また θ，ϕ，ϕ_l，ψ のそれぞれの確率分布は，最終的に得られたサンプルの集合から MAP 推定 [7] により得られる．

2.4 タグ付与

モデルの学習を終えた後，MAP 推定により得られた確率分布を用いて，タグの付与されていない文書に対してタグを付与する．

$$p(l|d,t) = \sum_z p(l|z)p(z|d,t_d) \cdot \frac{(1-t_d)^{\psi_{1,z}} \cdot t_d^{\psi_{2,z}}}{B(\psi_{1,z}, \psi_{2,z})} \tag{3}$$

$$\propto \sum_z p(l|z)p(z|d)p(z|t_d) \cdot \frac{(1-t_d)^{\psi_{1,z}} \cdot t_d^{\psi_{2,z}}}{B(\psi_{1,z}, \psi_{2,z})}$$

$$\propto \sum_z (\phi_l)^z_l \cdot (\theta)^d_z \cdot \frac{(1-t_d)^{\psi_{1,z}} \cdot t_d^{\psi_{2,z}}}{B(\psi_{1,z}, \psi_{2,z})} \tag{4}$$

付与するタグは，式 (4) により求められる $p(t|d)$ の値の大きいものから順に選ぶ．文書データに対する確率分布を得るために，モデルのパラメータ α，β，β_l，ϕ，ϕ_l，ψ を固定し，サンプリングにより $p(t|d)$ を得る．サンプリングの際に用いる timestamps は，文書 d が生成された時間 t_d を用いる．

タグ付与を行う過程を以下に示す．

学習データを提案モデルが学習する．学習で得たパラメータを保持し，未学習の文書データについて上述の通りパラメータを固定して学習し，$p(z|d)$ を得る．$p(z|d)$ の値にしたがって，上位 n 個のタグを選択し，タグ付与を行う．

3 実験

実際のレビュー文書を用いて，提案モデルの評価実験を行う．初めに，提案モデルの精度評価実験を行った．実験では，従来研究である Tag-LDA のタグ付与精度と F 値を用いて比較し，付与されるタグの精度について検討した．次に，提案モデルの時間変化に対するタグ付与性能の評価を行った．時間変化に対して付与されるタグについて，2 つのモデルで比較を行い，時間変化を考慮する妥当性を定性的に評価した．

3.1 実験条件

本実験では，So-net ブログ [3] に公開されているタグ付き文書のうち，2014 年の 1 年に書かれたブログ文書を用いた．その際，2014 年で話題となった文書を収集するために，前年度に比べて 2014 年に検索量が急上昇したワードのランキングを元に，ブログ内で検索をかけてデータの収集を行った．その中から名詞が 5 回以上 100 回以下出現し，全タグのうち，総出現回数が 4 回以上のタグに限定し，1082 件のブログ文書を収集した．

提案モデルにおける，ギブスサンプリングを用いた推論の反復回数は，学習データに対して 100 回，テ

[3] http://www.so-net.ne.jp/

ストデータに対して 100 回とした．提案モデルにおけるディリクレ分布のパラメータは，α=0.1，β=0.1，β_l=0.1 とし，トピック数は 10 とした．

3.2 提案モデルの精度評価実験

3.2.1 実験方法

文書へのタグ付与の精度評価について，従来研究との比較を行った．実験では，以下の試行を 1 試行とした．

まず，データセットの全ブログデータから 100 件をランダムに抽出してテストデータとし，残りを学習データとした．学習データを用いて各モデルの学習を行い，続いてテスト文書に対するトピックの確率分布推定した後，式（4）に基づいて 1~5 個のタグをテストデータに付与し，それぞれのタグ付与数における F 値を計算した．この操作を 10 回繰り返し，F 値の平均値と標準偏差を求めた．

以上の試行を 5 回繰り返し，F 値の平均値と標準偏差の平均値を計算した．

3.2.2 実験結果

従来研究の Tag-LDA と，提案モデルの F 値の平均値を標準偏差の平均値を図 2 に示す．図 2 から，提案モデルの F 値が，すべてのタグ付与数において劣っていることがわかる．タグ付与数が少ない時は Tag-LDA が大きく上回り，タグ付与数が多くなるほど，その差が小さくなる傾向がある．提案モデルにより実際に付与されたタグを確認してみると，一定期間に多く付与される話題のタグが適切に付与された文書があった一方で，多くの文書においては，同時期に多く付与されただけの，文書の内容とは異なるタグが付与されていた．話題の盛衰には合致しているものの，文書の内容を考慮できていないタグが付与されていたことから，今後はその 2 つを両立するような工夫が必要であると考えられる．

具体的には，提案モデルでは，学習データやテストデータで推論を行う際に，時間変化を考慮した推論を行うため，推定されるトピックも時間変化を考慮したものとなっている．また，タグ付与の際の式 (4) で，さらに timestamps を用いているため，時間変化を重視したタグ付与となってしまっている．そこで，推論またはタグ付与の際に，時間変化の要素を無くすこと，または，内容をより重視するような要素を加えることが改善につながると考えられる．

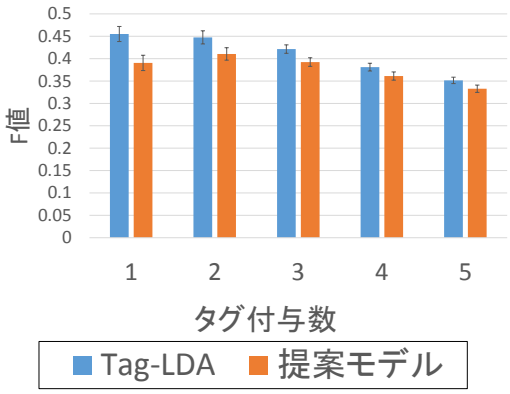

図 2: TagLDA と提案モデルの F 値による比較

3.3 時間変化に対するタグ付与性能の評価

3.3.1 実験方法

時間変化に対するタグ付与性能について，従来研究との比較を行った．実験では，前節でタグ付与数 5 の場合に付与されたタグを用いた．時間変化に対する，ある特定のタグの出現回数に着目し，話題の盛衰を捉えているかを定性的に評価した．

3.3.2 実験結果

タグ「全米オープン」「錦織圭」について，時間変化に対する，実際のタグ，テストデータでのタグ，Tag-LDA によるタグ，提案モデルによるタグをそれぞれ図 3(a)，図 3(b) に示す．

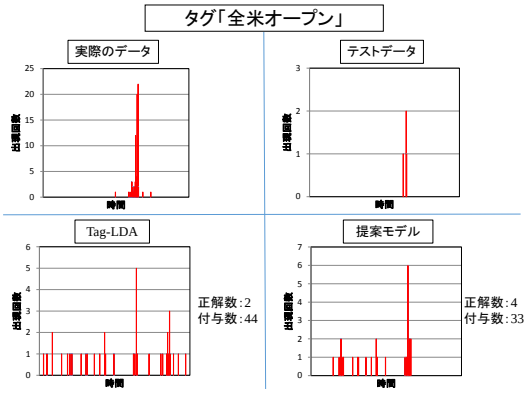

図 3(a): 全米オープン

まず，タグ「全米オープン」のように，ある一定期間に多く付与されているようなデータに対して，Tag-LDA は比較的どの時間でも同じようにタグを付与していることがわかる．一方，提案モデルでは実際のデータで多く付与されている時間，つまり話題となってい

る時期に多くのタグを付与している．また，Tag-LDA と比べて付与数が少ない一方で正解数が多く，わずかではあるが精度が向上していることは，時間変化を考慮することの効果であると思われる．

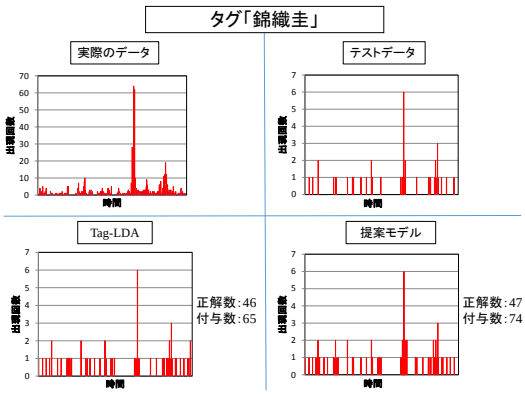

図 3(b): 錦織圭

次に，タグ「錦織圭」について，テストデータに対して，Tag-LDA は精度よく付与されている．一方で，提案モデルでは，実際のデータで多く付与されている時間に，多く付与していることがわかる．しかし，正解タグの付与数は 1 つ増えているものの，精度は Tag-LDA と比べて劣っている．実際に付与されたタグを確認すると，同時期には多く出現するものの，話題が異なる文書や，正解タグが「錦織圭」ではない文書に対して「錦織圭」のタグが付与されていた．内容を考慮できていないタグが付与されていたことから，前節と同様，今後は時間変化に加えて，さらに内容を考慮するような要素が必要であると考えられる．

4　おわりに

本稿では，時間とともに変化する話題を捉えた，タグ付与のためのトピックモデルを提案した．実際のレビュー文書を用いた実験により，F 値を用いて精度評価を行った．従来モデルである Tag-LDA と比較して，精度の面で提案モデルが劣っていることを確認した．また，特定のタグについて，時間変化に対する出現回数に着目し，話題の盛衰を捉えているかを定性的に評価した．一部のタグでは，時間変化と話題を捉えた所望のタグ付与が行われていたが，多くの文書において，時期的には捉えているが，話題が異なる文書に不正解となるタグ付与を行っているケースが見られた．今後は，モデルに対して内容をより重視するような要素を加えることが必要であると考えられる．

参考文献

[1] 西田京介, 藤村考: 階層的オートタギングによる Q & A コミュニティの知識整理, DEIM Forum, D3-4, 2010

[2] C.H. Brooks,and N. Montanez:Improved an- notation of the blogosphere via autotagging and hierar-chical clustering ,Proc.15thInternational Conference on World Wide Web, pp.625-632,2006

[3] S. Fujimura, K. Fujimura, and H. Okuda: Blogosonomy: Autotagging any text using blogger ’s knowledge, Proc. 2007 IEEE/WIC/ACM Interna-tional Conference on Web Intelligence,pp.205-212,2007

[4] T. Ohkura, Y. Kiyota, H. Nakagawa : Browsing System for Weblog Articles based on Automated Folksonomy, Proceedings of the WWW 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, at WWW 2006

[5] X. Si, M. Sun: Tag-LDA for scalable real-time tag recommendation, Joural of Computational Information Systems 6, pp. 23-31, 2009

[6] X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. In Proceedings of the 12th ACM SIGKDD international conference(KDD ’06), Philadelphia, PA, August 2006.

[7] Griffiths, T. L. and Steyvers, M.: Finding scientic topics, Proceedings of the National academy of Sciences of the United States of America, Vol. 101, No. Suppl 1, pp. 5228-5235 , 2004