トピックモデルに基づく文書ストリームのマルチラベル分類

白井　匡人†　　三浦　孝夫†

† 法政大学 工学研究科　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †masato.shirai.2d@stu.hosei.ac.jp, ††miurat@hosei.ac.jp

あらまし　本研究では文書ストリームを対象としたマルチラベル分類手法を提案する．文書ストリームでは新しい文書が逐次発生し文書集合が動的に変化することから，同一の分類基準を用いて分類を行うことは困難となる．このため，ラベルの特徴を動的に学習して分類を行う必要があるが，文書は一般に複数のラベルから構成されるため，マルチラベルでの特徴の変化も考慮することが求められる．提案手法では，トピックモデルによりラベルの特徴を動的に学習し，ラベル間の相関関係をラベリングに利用することで文書ストリームのマルチラベル分類を行う．

キーワード　文書ストリーム，トピックモデル，マルチラベル分類

1. 前　書　き

近年，インターネットの発達から大量の文書を入手することが容易になっている．これらの文書には多種多様な情報が含まれているが，未整理の大量の文書を人手で分類することは困難である．このため自動的に分類を行う手法が求められている．特に，ニュース記事，マイクロブログに代表されるストリームデータは極めて重要な情報源として注目されている．これらの文書集合ではリアルタイムな情報を扱うことから，ストリームデータの解析を行うことで現在の流行や評判情報など様々な知識を得られることが期待される．

次々に新たな文書が到着するという文書ストリームのような動的な文書集合では，話題の変化に応じて分類先となるクラスの特徴が変化することから，分類基準を固定することは困難となる．このようなストリームデータは新しいデータが逐次到着することからデータ量が膨大となり，集合で扱われる話題が動的に変化する．例えば，政治に関する話題では，選挙，予算編成，法案審議のように時期によるイベントや，外交問題や汚職事件など不定期なイベントに依存し，発生する単語分布は大きく異なる．これらのイベントの中でも，時間経過によって論じられる内容は逐次変化している．このため，文書ストリームの分類では，文書集合の変化を新たに到着した文書から学習し，分類基準を動的に変化させる適応的な分類手法が求められる．

また，ニュース記事やマイクロブログのようなストリームデータでは，季節や曜日などの周期的な要因や突発的なイベントなどの非周期的要因によって特定の話題に関する文書が大幅に増加する話題のバーストが発生する．話題のバーストとは特定のラベルに属する記事が短期的に急増する現象を指す．例えば，衆議院選挙のような大きなイベントが起これば，選挙期間中は他の話題に関する記事が減少する代わりに，衆議院選挙に関する記事が急増し，選挙が終わると異なる話題に遷移していき選挙に関する記事は減少する．このため，クラスの出現確率を考慮して分類を行うには，出現確率を固定して用いることはできず，高精度な分類を行うには，話題の出現の変化に応じてクラスの出現確率も動的に学習する必要がある．

文書分類では分類先となるラベルの特徴を学習することで分類を行うが，文書は一般に複数のラベルから構成されるため，マルチラベルでの特徴の変化も考慮する必要がある．文書のラベルには文書の話題に関連したラベルが付与されるため，共起するラベル間には相関関係があると考えられる．文書ストリームでラベルの相関関係を分類に用いるには，ラベル間の関係も動的に学習することが求められる．

本研究では文書ストリームを対象としたマルチラベル分類手法を提案する．提案手法では，トピックモデルの拡張により文書ストリーム上の，マルチラベルを持つ文書の特徴を動的に学習する．また，ラベル間の相関関係をラベリングに利用することで高精度に文書ストリームのマルチラベル分類を行えることを示す．

第 2 章では文書ストームについて述べ，第 3 章ではマルチラベル分類について述べる．第 4 章では提案手法について述べ，第 5 章では実験により有効性を示す．第 6 章で結論とする．

2. 文書ストリーム

2.1 ストリームの特性

ニュース記事，マイクロブログに代表されるストリームデータは近年極めて重要な情報源として注目されている．これらのデータは日々増加しており，出現する話題は逐次変化している．このような文書ストリームでは，文書は無限に増加するため，これまでに出現したすべての文書を用いて一括学習を行うことは困難である．このため，文書ストリームを扱うには，出現した文書を一度だけ学習に用いる単一パスで学習を行える差分的な手法が求められる．分類基準の変更が差分学習により精度を維持して実行可能であるならば，一括学習による再学習と比較して非常に高速であり，少ないメモリ量で分類が行える．

2.2 出現する話題の変化

ニュース記事のような文書ストリームでは，出現する話題が最近の出来事に大きく依存するため，特定の話題に関する記事が急増するバーストが起こることが知られている．例えば，図 1 より Reuters Corpus の 1996 年 8 月 20 日から 30 日間のラ

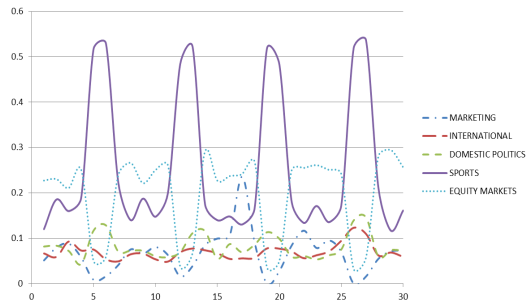


図 1　話題の出現確率

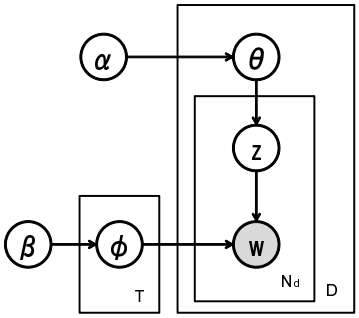
ベルの出現確率を見てみると，全体としては SPORTS の出現確率が高くなっているが，一定して出現確率が高くなっているわけではなく，EQUITY MARKETS と交互に最も出現確率が高いラベルとなっている．これらのラベルでは出現確率が周期的に変化している．MARKETING では出現確率が非周期的に変化しており，全体としては低い出現確率になっているが，ある時期では最も出現確率の高いラベルと同程度の確率となっている．このように，ラベルの出現確率は時期によって大きく変化しており，ラベルの出現確率 $P(c)$ を分類に用いるには，ラベルの単語分布と同様に変化に応じて更新する必要がある．また，ラベルの出現確率は季節や曜日などの周期的要素や大きなイベントなど非周期的な要素に依存するため，ラベルの出現確率が急増したラベルでは新たな話題が発生している可能性がある．このため文書ストリームを扱うには，動的に学習を行いラベルの特徴の変化を捉える必要がある．

3.　マルチラベル分類

マルチラベル分類では文書が複数のラベルに属するという仮定から，文書の持つラベルベクトル $\mathbf{y_i} = y_{i,1}, y_{i,2}, ..., y_{i,n}$ を推定する．ここで y_i は $y_i \in 0, 1$ であり，属するラベルの値が 1，属さないラベルが 0 のラベルベクトルである．単一ラベルの分類ではラベルごとに学習を行うが，マルチラベルでは組み合わせ数が $2^l - 1$ と膨大になるため，ラベルの組み合わせであるラベルセットごとに学習を行うことは困難となる．このため，複数のラベルを持つ文書から単一ラベルの特徴を学習して分類を行う必要がある．

学習されたラベルを基に文書のラベルベクトルの推定を行うが，尤度の最も高くなるラベルベクトルを導出するのにも学習と同じく $2^l - 1$ の尤度の計算が必要となるため，サンプリングや事後確率が改善されなくなるまで最大尤度のラベルを追加するなどの近似を行うことで，ラベルベクトルの推定が行われている [4]．しかしながら，これらの手法ではラベル推定時にラベルの相関関係を考慮していないため，ラベルの共起等の出現しやすいラベルの組み合わせを推定することが困難となる．ラベル間には”経済”と”市場”は共起しやすいが”芸術”と”健康”は共起しにくい等の相関関係があることが考えられる．

また，文書ストリームのマルチラベル分類では，ラベルの特徴を動的に学習すると共にラベル間の相関関係についても学習を行う必要がある．文書ストリームではラベルの出現確率が大きく変動するため，ラベル間の関係も変化していくことが考えられる．本研究では，ラベル間の関係を動的学習することでストリームの変化に対応した分類を行う．



(a)

図 2　LDA

4.　トピックモデル

トピックモデルは文書集合の分析に非常に有効なモデルとして注目されており，文書分類 [9] や文書要約など幅広く用いられている．トピックモデルとは, 1 つの文書が複数のトピックの混合として表現されるという仮定である．1 つの文書が 1 つのトピックで表される混合多項分布に比べ, トピックモデルは文書が複数のトピックの混合分布として, 各トピックが単語の分布として表現され, 高い精度で文書をモデル化する可能性がある．その中でも最近用いられているのが LDA である．確率的潜在意味索引付け (Probablisitic Latent Semantic Indexing, pLSI) は, LDA と違って, トピックと単語の多項分布にそれぞれディリクレ事前分布を導入し, 混合比を学習データの文書集合に依存して固定化している．

一方, LDA ではこの混合比は事前分布から動的に生成する点で異なる．LDA は学習データだけに依存しないが, 代わりに特定の確率モデルを仮定するため, 柔軟な観点で文書を扱える可能性がある．

図 2 は LDA のグラフィカルモデルである．図中の変数は, 図 2 左に, ディリクレ事前分布 $Dir(\beta)$, 図 2 左下の単語空間の多項分布 $Multinomial(\phi_{z_i})$, T はトピック数, 図 2 上にディリクレ事前分布 $Dir(\alpha)$, 図 2 中央にトピック空間の多項分布 $Multinomial(\theta_d)$, D は文書数, N は各文書の単語数を表す．LDA の単語生成過程を以下で示す．まず, すべてのトピック t においてディリクレ事前分布 $Dir(\beta)$ から ϕ_t を抽出し, 同様に, すべての文書 d においてもディリクレ事前分布 $Dir(\alpha)$ から θ_d を抽出する．次に, 文書 d 内の i 番目の単語 w_i において，抽出した文書 d の多項分布 $Multinomial(\theta_d)$ からトピック z_i を抽出し，そのトピック z_i の多項分布 $Multinomial(\phi_{z_i})$ から単語 w_i を抽出する．

5. 提案手法

5.1 提案モデル

本研究ではマルチラベルを持つ文書集合の解析を行う手法として，Multi-label topic model を提案する．提案モデルでは文書の持つラベルベクトルと語集合から，単語を生成する潜在変数であるラベルとトピックを推定することで文書集合のトピック分布を学習する．提案モデルのグラフィカルモデルを図 3 に示す．図中の変数は, 図 3 左に, ディリクレ事前分布 $Dir(\alpha)$・$Dir(\beta)$, 図 3 左上のトピック空間の多項分布 $Multinomial(\theta_l)$，L はラベル数，図 3 左下の単語空間の多項分布 $Multinomial(\phi_{z_i})$, T はトピック数, 図 3 上にディリクレ事前分布 $Dir(\gamma)$, 図 3 中央にラベル空間の多項分布 $Multinomial(\lambda_j)$, 文書 d のラベルセット a，D は文書数, N は各文書の単語数を表す．提案モデルの単語生成過程を以下で示す. まず, ディリクレ事前分布 $Dir(\gamma)$ からラベルセット a に含まれる λ_jを抽出し，文書 d 内の i 番目の単語 w_i のラベル l を選択する．次いで，選択されたのラベルのトピック t においてディリクレ事前分布 $Dir(\beta)$ からトピックの語分布である ϕ_t を抽出し，ディリクレ事前分布 $Dir(\alpha)$ からラベルのトピック分布である θ_l を抽出する．抽出したラベル j の多項分布 $Multinomial(\theta_l)$ からトピック z_i を抽出し，そのトピック z_i の多項分布 $Multinomial(\phi_{z_i})$ から単語 w_i を抽出する．ただしラベルの持つトピックは独立しており，ラベル数 L，各トピック数が M であるとき，トピックベクトルは $T = t_{1,1}, t_{1,2}, ...t_{1,m}, t_{2,1}, ..., tl, m$ で表される. 文書 d の単語 i のラベルがラベル k となる確率はギブスサンプリングにより以下の式で近似される．

$$p(x_i = k|w_i = m, \lambda_{-i}, \mathbf{a}, \mathbf{x}_{-i}, \mathbf{w}_{-i}) \propto \frac{C^{WL}_{mk} + \beta}{\sum_m C^{WL}_{mk} + V\beta} \frac{C^{LD}_{kD} + \gamma}{\sum_k C^{LD}_{kd} + A\gamma} \tag{1}$$

ここで C^{WL}_{mk} はラベル k が単語 m に割り当てられた回数，C^{LD}_{kD} は文書 d にラベル k が割り当てられた回数，V は全単語数，A は文書 d のラベル数である．また，文書 d の単語 i のラベルが k，トピックが j となる確率は以下の式で近似される．

$$p(z_i = j, x_i = k|w_i = m, \mathbf{z}_{-i}, \lambda_{-i}, \mathbf{a}, \mathbf{x}_{-i}, \mathbf{w}_{-i}) \propto \frac{C^{WT}_{mj} + \beta}{\sum_m C^{WT}_{mj} + V\beta} \frac{C^{LT}_{kj} + \alpha}{\sum_j C^{LT}_{kj} + L\alpha} \tag{2}$$

C^{WT}_{mj} はトピック j が単語 m に割り当てられた回数，C^{LT}_{kj} はラベル k にトピック j が割り当てられた回数，T は全トピック数である．またトピックの単語分布 ϕ とラベルのトピック分布 θ は以下の式で求まる．

$$\phi_{mj} = \frac{C^{WT}_{mj} + \beta}{\sum_m C^{WT}_{mj} + V\beta} \tag{3}$$

$$\theta_{kj} = \frac{C^{LT}_{kj} + \alpha}{\sum_j C^{DT}_{dj} + T\alpha} \tag{4}$$

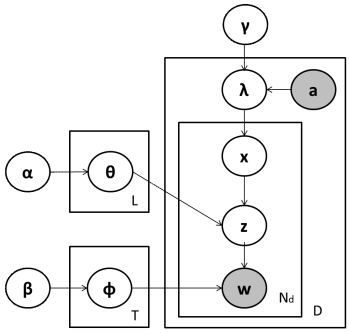


図 3　提案モデル

5.2 ラベリング手法

未知の文書のラベルベクトルの推定には学習されたパラメータとラベルの共起確率を用いる．ラベリングではまず各ラベルから未知の文書が生成される尤度を算出し，最も尤度の高くなる単一ラベル y_i を決定する．次いで決定された単一ラベルを基に共起するマルチラベルのセットを求める．最後にラベルセットの中で最も尤度の高くなるマルチラベルをラベルベクトルの推定結果とする．また，本研究では文書ストリームを対象としているためウィンドウを設け特定の期間の中でラベルセットと共起確率を求める．ウィンドウはサイズ S の直近の文書集合であり現在の文書 d_t から $d_t, dt-1, ..., dt-s$ で表される．これにより，分類先となるラベルセットを動的に変更しながらラベリングを行う．ラベリングの手順と尤度の計算式を以下に示す．

（1） ラベル集合 $y_1, ..., y_n$ から尤度が最大となるラベル y_i を求める．

（2） ウィンドウ S 内でラベル y_i と共起するすべてのラベルセットを求める．

（3） ラベルセットの中から尤度が最大となるラベルを求める．

$$Ans(l) = arg \ \max_l P(l)P(d|\phi_l, \theta_l) \tag{5}$$

ここで，$P(l)$ はクラスの出現確率，$P(d|\phi_l,\theta_l)$ は各ラベルから未知の文書が生成される尤度である．

6. 実　　験

本章では Reuters Corpus を用いてマルチラベル分類を行う．比較手法として提案手法で用いるパラメータ更新，共起確率，ウィンドウをそれぞれ使用しない分類を行い，提案手法の有効性を検証する．

6.1 実験準備

実験に用いる Reuters Corpus は 1996 年 8 月 20 日から 1997 年 8 月 19 日までの 1 年分のニュース記事であり，1 つの記事に 128 種類からなるラベルが複数付いている．本実験では，頻度が上位となるマルチラベルに含まれるラベルを頻度が高い順に 20 個抽出して用いる．用いるラベルと各ラベルを含む文書数を表 1 に示す．実験データには 8 月 20 日から 11 月 30 日までの 103 日分を用いる．総文書数は 103925 であり，このうち先頭から 3000 文書を学習データとして用いる．

トピックモデルの各パラメータの値はトピック数 10，ギブスサンプリングの繰り返し回数 300 回とし，事前分布のパラメータは α=0.5，β=0.01，γ=0.01 とする．インクリメンタルギブスサンプリングの繰り返し回数は 20 回とする．また，テスト文書の 10% を教師データとして分類実行後に正解を与えて学習に使用する．ラベルの共起を取るためのウィンドウサイズは 3000 文書とする．

6.2 評価方法

実験の評価にはラベル単位と文書単位の f 値を用いる．f 値は再現率と適合率の調和平均であり，再現率は実際に正であるもののうち，正であると予測されたものの割合，適合率は正と予測したデータのうち，実際に正であるものの割合である．ラベル単位の f 値を以下の式で求める．

$$R_i = \frac{a_i}{a_i + c_i} \tag{6}$$

$$P_i = \frac{a_i}{a_i + b_i} \tag{7}$$

a_i は推定結果が正である数，c_i は正であるが負と推定された数，b_i は正であると推定した中で正解が負である数である．この 2 つの式の調和平均である f 値を次のように定義する

$$f_i = \frac{2 \times P_i \times R_i}{P_i + R_i}, \quad f = Average_i f_i$$

また，文書単位の f 値を正解のラベルベクトル $y_i = (y_{i,1}, ..., y_{i,n})$ と予測したラベルベクトル $y_i^{'} = (y_{i,1}^{'}, ..., y_{i,n}^{'})$ 以下の式で求める．

$$R_i = \frac{\sum_{l=1}^{L} y_{n,l} y_{n,l}^{'}}{\sum_{l=1}^{L} y_{n,l}} \tag{8}$$

$$P_i = \frac{\sum_{l=1}^{L} y_{n,l} y_{n,l}^{'}}{\sum_{l=1}^{L} y_{n,l}^{'}} \tag{9}$$

$$F_i = \frac{2 P_i R_i}{P_i + Ri} \tag{10}$$

6.3 実験結果

実験結果を表 2 に示す．F 値は更新無しのクラス単位で 0.632，文書単位で 0.718，共起確率を考慮しない手法でクラス単位 0.665，文書単位 0.735，ウィンドウを設けない手法でクラス単位 0.691，文書単位 0.730 であり，パラメータ更新・共起確率・ウィンドウのすべてを用いた提案手法ではクラス単位 0.702，提案手法で 0.776 となっている．

7. 考　　察

表 2 より，パラメータ更新，共起確率，ウィンドウを用いた提案手法が最も精度が高くなっている．まずパラメータ更新の効果としてストリーム中の新たな教師データから学習を行うことで F 値がクラス単位で 0.07，文書単位で 0.058 上昇している．このことから提案モデルのインクリメンタルギブスサンプリングを用いた差分学習によって適切な学習が行えていると考えられる．次に共起確率の効果として最大尤度となる単一ラベルのラベルセットに対して，共起確率による重み付けを行うことで F 値がクラス単位で 0.037，文書単位 0.041 上昇している．これはマルチラベルとなるラベル間には相関関係が有り，ラベリング時に用いることで精度が上昇したと考えられる．最後にウィンドウの効果として，ラベル間の共起確率をウィンドウ内で求めることで学習データ全体を通した共起確率と比較して F 値がクラス単位で 0.011，文書単位で 0.046 上昇している．これはラベルの出現確率が周期的或いは非周期的な要因によって大きく変化するというストリームの特性を捉え，ウィンドウによって直近の期間の変動を考慮してラベリングを行えているためであると考えられる．提案手法ではクラス単位の F 値と文書単位の F 値がすべて比較手法と比して上昇しているが，特にラベリングに関する手法である共起確率とウィンドウでは文書単位の F 値がクラス単位の F 値より大きく上昇している．これはラベリング手法が単一ラベルからラベルセットの推定を行うマルチラベル分類に対して有効に機能しているためであると考えられる．

また，提案手法の F 値はクラス単位で 0.702，文書単位で 0.776 と高精度になっていることから提案モデルによりマルチラベルを持った文書から適切な学習を行えていると考えられる．提案手法ではラベリング時にラベルの共起関係とウィンドウを用いることで分類先となるラベルセットを動的に変更しながら分類を行ったが，モデルの学習に用いるラベル数は固定している．ストリームでは時間に応じて新たな話題が発生することから，ラベル自体が新たに生成することも考えられる．また，あるラベルで発生する話題が大きく変化し一つのラベルとして扱えなくなることや複数のラベルを一つのラベルとして扱うことで柔軟にモデル化を行える可能性がある．ラベリング時だけでなく，モデルの学習時にもクラスの分化，結合や生成を行い，ラベル数自体を動的に変更しながら分類を行うことで精度が上昇することが期待される．

8. 結　　論

本研究では文書ストリームに対するマルチラベル分類を行う手法を提案した．マルチラベル分類の結果，F 値はクラス単位で 0.702，文書単位で 0.776 と高精度で行えることを示した．また，トピックモデルの動的学習を行い，ストリーム上のマルチラベルを持つ文書の特徴を学習することで F 値がそれぞれ 0.07,0.058 上昇し，ラベルの共起関係を用いることで 0.037，0.041，ウィンドウを用いることで 0.011，0.046 上昇することを示した．

文　　献

[1] Hofmann, T.: Probabilistic Latent Semantic Indexing, SIGIR, 1999

[2] Blei, D. M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, Journal of Machine Learning Research 3, pp. 993-1022, 2003.

[3] Rosen-Zvi, M., Griffiths, Steyvers, M. and Smyth, T.: The author-topic model for authors and documents *UAI '04 Proceedings of the 20th conference on Uncertainty in artificial intelligence* , 2004.

[4] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D.Manning "Labeled LDA:A supervised topic model for credit attribution in multi-labeled cor-

	C15	C151	C152T	C18	C181	CCAT	E21	E212	ECAT	GCAT
文書数	36489	19233	18556	7358	7604	42985	5472	5444	5509	15671
	GSPO	M11	M12	M13	M131	M132	M14	M141	M143	MCAT
文書数	9073	10760	5644	10557	6702	4358	14659	8979	4545	38921

表 1　実験データ

	C15	C151	C152T	C18	C181	CCAT	E21	E212	ECAT	GCAT	GSPO
(更新無し)											
再現率	0.822	0.851	0.912	0.870	0.929	0.929	0.782	0.721	0.577	0.749	0.746
適合率	0.688	0.711	0.699	0.684	0.645	0.457	0.208	0.606	0.293	0.241	0.248
F 値	0.749	0.775	0.792	0.766	0.762	0.612	0.328	0.659	0.388	0.364	0.372
(重み無し)											
再現率	0.676	0.863	0.764	0.850	0.890	0.962	0.754	0.897	0.876	0.813	0.809
適合率	0.733	0.729	0.800	0.856	0.791	0.461	0.296	0.476	0.250	0.325	0.315
F 値	0.703	0.790	0.781	0.853	0.838	0.623	0.425	0.622	0.389	0.464	0.454
(ウィンドウ無し)											
再現率	0.687	0.854	0.792	0.848	0.864	0.954	0.757	0.872	0.797	0.806	0.801
適合率	0.748	0.756	0.784	0.857	0.796	0.551	0.334	0.561	0.312	0.381	0.358
F 値	0.716	0.802	0.788	0.853	0.828	0.699	0.463	0.683	0.448	0.518	0.495
(提案手法)											
再現率	0.705	0.851	0.855	0.791	0.890	0.961	0.729	0.886	0.811	0.797	0.793
適合率	0.748	0.754	0.895	0.811	0.810	0.515	0.347	0.552	0.317	0.436	0.417
F 値	0.726	0.800	0.875	0.801	0.848	0.670	0.470	0.681	0.455	0.564	0.547
	M11	M12	M13	M131	M132	M14	M141	M143	MCAT	クラス単位	文書単位
	0.951	0.616	0.904	0.906	0.901	0.738	0.980	0.972	0.625	0.824	0.862
	0.641	0.342	0.279	0.278	0.280	0.940	0.855	0.512	0.654	0.513	0.615
	0.766	0.440	0.426	0.426	0.427	0.827	0.913	0.671	0.639	0.632	0.718
	0.923	0.699	0.844	0.843	0.857	0.820	0.941	0.986	0.799	0.843	0.837
	0.413	0.445	0.437	0.440	0.419	0.774	0.824	0.795	0.391	0.549	0.655
	0.571	0.544	0.576	0.578	0.563	0.796	0.879	0.880	0.525	0.665	0.735
	0.945	0.658	0.882	0.882	0.815	0.828	0.910	0.987	0.818	0.838	0.833
	0.516	0.464	0.465	0.463	0.451	0.885	0.821	0.785	0.486	0.589	0.711
	0.667	0.544	0.609	0.607	0.581	0.856	0.863	0.874	0.610	0.691	0.730
	0.922	0.669	0.845	0.846	0.848	0.832	0.946	0.988	0.790	0.838	0.841
	0.561	0.470	0.510	0.508	0.505	0.846	0.827	0.782	0.482	0.605	0.721
	0.697	0.552	0.636	0.635	0.633	0.839	0.882	0.873	0.599	0.702	0.776

表 2　マルチラベル分類結果

pora”,Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing(EMNLP’09),pp.248-256,2009.

[5] Yao, L., Mimno D., McCallum, A.: Efficient Methods for Topic Model Inference on Streaming Document Collections, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.937-946, 2009

[6] Canini, K.R., Shi, L., Griffiths, T.L.: Online Inference of Topics with Latent Dirichlet Allocation, Journal of Machine Learning Research, Proceedings Track 5, pp.65-72, 2009

[7] Sato, I., Kurihara, K., Nakagawa, H.: Deterministic Single-Pass Algorithm for LDA, proc. 24th Annual Conference on Neural Information Processing Systems (NIPS), 2074-2082, 2010

[8] Hoffman, M.D., Blei, D.M., Bach, F.R.: Online Learning for Latent Dirichlet Allocation, proc. 24th Annual Conference on Neural Information Processing Systems (NIPS), pp.856-864, 2010

[9] Masato Shirai, and Takao Miura “On Domain Independence of Author Identification”,IDEAL’11, 2011