DEIM Forum 2017 A3-5

潜在変数モデリングによる利用者の情報行動に関する予測

梅原　頌平†　　江口　浩二††

† 神戸大学システム情報学研究科　〒 657–8501 神戸市灘区六甲台町 1-1
†† 神戸大学システム情報学研究科　〒 657–8501 神戸市灘区六甲台町 1-1
E-mail: †umehara@cs25.scitec.kobe-u.ac.jp, ††eguchi@port.kobe-u.ac.jp

あらまし　オブジェクト間の関係を表す関係データはグラフとして表現することができ，そのときグラフにおける頂点はオブジェクト，辺は関係に対応する．また，複数種類の関係からなる多次元関係データは，複数種類の辺からなる多モードネットワークとして表現できる．ウェブ検索クエリログにおいてもクエリ間の複雑な関係を多次元関係データあるいは多モードネットワークと見なすことができる．以上の問題を解決するために，本論文では新たに多モードブロックモデル（multi-mode block model）を提案する．そして，本論文では現実の検索クエリログのデータを用いた実験により，多モードブロックモデルの有効性について評価を行う．

キーワード　検索行動分析, クエリ提案, クエリログ, 潜在変数モデル, ブロックモデル

1. はじめに

オブジェクト間の関係を表す関係データはグラフとして表現することができ，そのときグラフにおける頂点はオブジェクト，辺は関係に対応する．また，複数種類の関係からなる多次元関係データは，複数種類の辺からなる多モードネットワークとして表現できる．多次元関係データの例は枚挙に暇がないが，共通の趣味を持つ人間関係や共通の出身校である人間関係などからなるソーシャルネットワークが典型的である． ウェブ検索クエリログにおいても共通のウェブページに対してクリックが行われたクエリの対や，同一の検索セッションにおいて投入されたクエリの対，部分的に共通した文字列からなるクエリの対などのような，クエリ間の複雑な関係を多次元関係データあるいは多モードネットワークと見なすことができる．

さて，種々の離散データ集合の分析手段の一つとして混合多項分布モデル（Multinomial mixture models) [1] やそれを拡張したトピックモデル（Topic models）[2] が有効であることが知られている．文書データ集合を例にとって説明すると，混合多項分布モデルは文書毎に単語の分布として表現される潜在的なトピックが背後に存在すると仮定するモデルである．ここでは文書毎に一つの潜在トピックが対応付けられるのに対して，トピックモデルは各文書を複数の潜在トピックの混合分布として表現するモデルである．トピックモデルの代表的なものとしては潜在ディレクレ配分法（Latent Dirichlet Allocation: LDA）[3] が挙げられる．混合分布モデルやトピックモデルの適用範囲は広く，その拡張形が関係データやネットワークの分析にも適用されている．

関係データまたはネットワークを扱えるトピックモデルの一つとして，短い長さのテキストデータにおける単語間の関係に着目したバイタームトピックモデル（Biterm topic model: BTM）[4] が挙げられる．バイタームトピックモデルでは関係データにおける各オブジェクト対について共通のトピックが仮定されている．それに対して，関係データにおける各オブジェクト対について異なるトピックを仮定するスパースブロックモデル（Sparse block model: SBM）[5] がある．さらには，この SBM と先に述べた LDA を統合し，関係データと文書データを横断してモデリングするブロックトピックモデル（Block-LDA）[6] が開発されている．これらのモデルは均質な関係データやネットワークを想定したものであり，多次元関係データや多モードネットワークにおける種々の関係の不均質性を捉えることはできない．

以上の問題を解決するために，本論文では新たにトライモード・ブロックモデル（Tri-mode block model）を提案する．このモデルは多モード，特に 3 種類のモードで構成されるネットワークをモデリングするもので，3 種類のモードが共通のトピックの分布から生成されると考えるものである．本論文では現実の検索クエリログのデータを用いた実験により，トライモード・ブロックモデルの有効性を評価する．

2. 関連研究

ここでは提案手法に関連した研究として，混合多項分布モデル及びバイタームトピックモデル，スパースブロックモデル，ブロックトピックモデルについて説明する．

2.1 混合多項分布モデル

混合多項分布モデル (Multinomial of mixture models) は，文書毎に単語の分布として表現される潜在的なトピックが背後に存在すると仮定するモデルである．すなわち各文書はそれぞれ 1 つのトピックで表現され，文書内の単語はそのトピックから生成される．またトピックはそれぞれ異なった単語分布として表現される．混合多項分布モデルのグラフィカルモデルを図 1 に示す．図中の D, N_d, K はそれぞれ文書数，文書 d の単語数，トピック数である．θ, ϕ_k はそれぞれトピックの多項分布パラメータ，トピック k に関するデータの多項分布パラメータである．α, β はそれぞれ上記 2 種類の多項分布に対応するディレクレハイパーパラメータである．混合ユニグラムモデルにおける文書の生成過程を以下に示す．

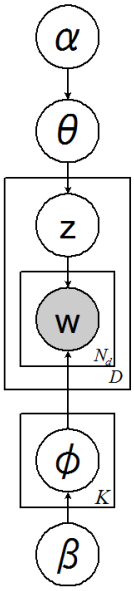

図 1　Grephical model representation of multinomial mixture model.

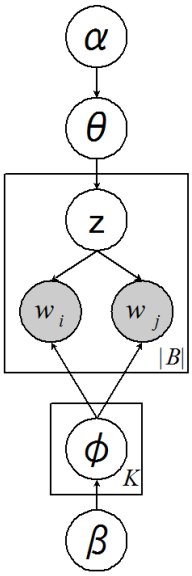

図 2　Grephical model representation of BTM.

（1） 文書全体に対し，$\theta \sim Dirichlet(\alpha)$ を選択する．

（2） K 個のトピックに対し，$\phi_k \sim Dirichlet(\beta)$ を選択する．

（3） 文書 d に対し:

a　トピック $z_d \sim Multinomial(\theta)$ を選択する．

b　N_d 個の単語 w_{di} に対し，単語 $w_{di} \sim Multinomial(\phi_{z_d})$ を選択する．

ここで $Dirichlet$ はディレクレ分布，$Multinomial$ は多項分布を表している．周辺化ギブスサンプリングの完全条件付確率は以下の式で与えられる:

$$p(z_d = k \mid \mathbf{W}, \mathbf{z}^{\neg d}, \alpha, \beta) \propto (D_k^{\neg d} + \alpha) \cdot \frac{\Gamma(N_k^{\neg d} + \beta V)}{\Gamma(N_k^{\neg d} + N_d + \beta V)} \prod_{v:N_{dv}>0} \frac{\Gamma(N_{kv}^{\neg d} + N_{dv} + \beta)}{\Gamma(N_{kv}^{\neg d} + \beta)} \quad (1)$$

ここで $\mathbf{w} = \{w_{di}\}$，$\mathbf{z} = \{z_{di}\}$，上付き文字「$\neg d$」は文書 d の要素をデータから除くことを示している．V は語彙数を表す．また，D_k はトピック k が割り当てられた文書数，N_k はトピック k が割り当てられた文書における総単語数，N_{kv} はトピック k が割り当てられた文書における語彙 v の出現回数を表す．

2.2　バイタームトピックモデル

バイタームトピックモデル (Biterm Topic Model: BTM) は，短い長さのテキストデータに対して単語共起性に着目したトピックモデルである．従来のトピックモデル (LDA や PLSA) は文書単位の単語共起性を捉えるものであるため，ショートテキストでは単語共起性が疎になるという問題を抱えている．そこで BTM はコーパス全体の単語共起性を捉えることでこの問題を解決しようとしている．BTM では関係データにおける各オブジェクト対を「バイターム」として抽出して考える．BTM の特徴として，同じバイタームの 2 つのオブジェクトは同じトピックに属すると仮定されている．BTM のグラフィカルモデルを図 2 に示す．$\mathbf{B}$ はバイタームの集合を表し，図中の $|\mathbf{B}|$ はその総数を表す．BTM における文書の生成過程を以下に示す．

（1） バイターム集合全体に対し，$\theta \sim Dirichlet(\alpha)$ を選択する．

（2） K 個のトピックに対し，$\phi_k \sim Dirichlet(\beta)$ を選択する．

（3） バイターム b に対し:

a　トピック $z_b \sim Multinomial(\theta)$ を選択する．

b　バイタームの 2 単語 w_{bi}, w_{bj} に対し，単語 $w_{bi}, w_{bj} \sim Multinomial(\phi_{z_b})$ を選択する．

周辺化ギブスサンプリングの完全条件付き確率は以下の式で与えられる:

$$p(z = k \mid \mathbf{B}, \mathbf{z}^{\neg b}, \alpha, \beta) \propto (B_k^{\neg b} + \alpha) \frac{(N_{v_i k}^{\neg b} + \beta)(N_{v_j k}^{\neg b} + \beta)}{(\sum_v N_{vk}^{\neg b} + |\mathbf{B}|\beta)^2} \quad (2)$$

ここで，B_k はトピック k に割り当てられたバイタームの総数，N_{vk} は語彙 v がトピック k に割り当てられた回数，上付き文字「$\neg b$」はバイターム b の要素をデータから除くことを示している．

2.3　スパースブロックモデル

スパースブロックモデルは，タンパク質の相互作用やソーシャルネットワークの分析などの関係データ間のリンクをモデリングするブロックモデルである．関係データにおける各オブジェクト対についてそれぞれ異なるトピックを仮定していることが 2.2 節の BTM との違いである．スパースブロックモデルのグラフィカルモデルを図 3 に示す．$\mathbf{L}$ はリンクの集合を表し，図中の $|\mathbf{L}|$ はその総数を表す．スパースブロックモデルにおける生成過程を以下に示す．

（1） リンク全体に対し，$\theta \sim Dirichlet(\alpha)$ を選択する．

（2） K 個のトピックに対し，$\phi_k \sim Dirichlet(\beta)$ を選択

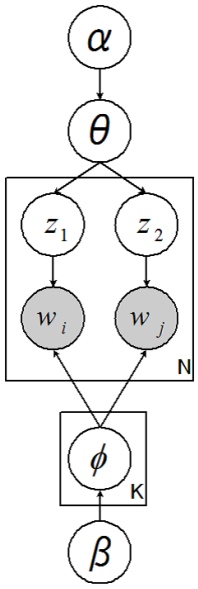

図 3　Grephical model representation of Sparse Block Model.

する．

（3） リンク $\ell = (i, j)$ に対し:

a　トピックペア $z_\ell \sim Multinomial(\theta)$ を選択する．

b　リンク ℓ の両端の 2 つのエンティティ w_{li}, w_{lj} に対し，

エンティティ $w_{li} \sim Multinomial(\phi_{z_1}), w_{lj} \sim Multinomial(\phi_{z_2})$ をそれぞれ選択する．

ここで $z_\ell = (z_1, z_2)$ であり，単一のトピックではなくトピックの対を指しており，それぞれのトピックからエンティティが選択される．また，$\mathbf{z} = \{z\}_\ell$ である．周辺化ギブスサンプリングの完全条件付き確率は以下の式で与えられる:

$$p(z_\ell \mid \mathbf{z}^{\neg\ell}, \mathbf{L}^{\neg\ell}, \alpha, \beta) \propto (D_k^{\neg\ell} + \alpha) \cdot \frac{(N_{k_1 i}^{\neg\ell} + \beta)(N_{k_2 j}^{\neg\ell} + \beta)}{(N_{k_1 \cdot}^{\neg\ell} + N\beta)(N_{k_2 \cdot}^{\neg\ell} + N\beta + \delta_k)} \quad (3)$$

ここで上付き文字「$\neg\ell$」はリンク ℓ を除くという事を示し，D_k はトピックペア k に割り当てられた回数，N_{ki} は単語 i がトピック k に割り当てられた回数，$N_{k\cdot}$ はトピック k が割り当てられた回数，δ_k は $k = (k_1, k_2)$ で $k_1 = k_2$ の時 1，それ以外の時 0 を示す．

2.4　ブロックトピックモデル

ブロックトピックモデル (Block-LDA) は，エンティティ間のリンクとリンクされたエンティティに関する文書を連結してモデリングすることでエンティティ間のリンクモデリングを改善することを目的としたモデルである．このモデルでは確率論的ブロックモデルにおけるブロックとトピックモデルにおける潜在トピックのアイデアを連結している．ブロックトピックモデルのグラフィカルモデルを図 4 に示す．図の左側がリンクモデリングであり 2.3 節で述べたスパースブロックモデルが元になっている．また，図中の α_L は文書のトピック多項分布に関するハイパーパラメータ，α_D はリンクのトピック多項分布に関するハイパーパラメータを表す．一方，図の右側は関連文書のモデリングであり潜在的ディレクレ配分法 (Latent DIrichlet Allocation:LDA) が元になっている．ブロックトピックモデルにおける文書の生成過程を以下に示す．

（1） T 種類毎の K 個のトピックに対し，$\phi_{t,z} \sim Dirichlet(\beta)$ を選択する．

（2） 文書 d に対し:

a　文書トピック分布 $\theta_D \sim Dirichlet(\alpha_D)$ を選択する．

b　タイプ t のエンティティ集合 $w_{t,i}$ に対し:

- トピック $z_{t,i} \sim Multinomial(\theta_d)$ を選択する．
- エンティティ $w_{t,i} \sim Multinomial(\phi_{t,z_{t,i}})$ を選択する．

（3） タイプ t_ℓ のエンティティのリンクに対し:

a　トピックペア分布 $\theta_L \sim Dirichlet(\alpha_L)$ を選択する．

b　リンク $w_{i1} \to w_{i2}$ に対し:

- トピックペア $(z_{i1}, z_{i2}) \sim Multinomial(\theta_L)$ を選択する．
- リンク $w_{i1} \sim Multinomial(\phi_{t_l, z_{i1}})$ を選択する．
- リンク $w_{i2} \sim Multinomial(\phi_{t_l, z_{i2}})$ を選択する．

文書内のエンティティの周辺化ギブスサンプリングの完全条件付確率は以下の式で与えられる:

$$p(z_{t,i} = k \mid \mathbf{Z}^{\neg i}, \mathbf{W}_t^{\neg i}, \alpha_D, \beta) \propto (N_{dk}^{\neg i} + \alpha_D) \frac{N_{ktv_{t,i}}^{\neg i} + \beta}{\sum_{v'} N_{ktv'}^{\neg i} + |\mathbf{W}_t|\beta} \quad (4)$$

ここで上付き文字「$\neg i$」はリンク i を除くという事を示し，N_{dk} は文書 d でトピック k が割り当てられた回数，N_{ktv} は語彙 v がタイプ t のトピック k に割り当てられた回数を表す．

一方，リンクの周辺化ギブスサンプリングの完全条件付確率は以下の式で与えられる:

$$p(z_i = (k_1, k_2) | (v_{i1}, v_{i2}), z^{\neg i}, (v_1, v_2)^{\neg i}, \alpha_L, \beta) \propto (L_{(k_1,k_2)}^{\neg i} + \alpha_L) \times \frac{(N_{k_1 t_\ell v_{i1}}^{\neg i} + \beta)(N_{k_2 t_\ell v_{i2}}^{\neg i} + \beta)}{(\sum_v N_{k_1 t_\ell v}^{\neg i} + |\mathbf{W}_{t_\ell}|\beta)(\sum_v N_{k_2 t_\ell v}^{\neg i} + |W_{t_\ell}|\beta)} \quad (5)$$

ここで $L_{(k_1,k_2)}$ はトピックペア (k_1, k_2) が割り当てられたリンクの総数を表す．β が (4) 式と同一であり，これによりネットワークと文書を結合している事に注意されたい．

3.　提 案 手 法

3.1　多モード・ブロックモデル (Multi-mode block model)

ネットワークにおいて辺が二種類の集合に分かれるようなネットワークを 2 モードネットワーク集合と呼ぶ．この 2 モードネットワークを一般化し，均質でない辺からなるネットワークを多モードネットワークとする．図 5 は 2 モードネットワークの例を示し，図 6 は多モードネットワークの例を示す．2 モードネットワークでは 2 種類の辺に分かれており，多モードネットワークでは複数種類 (図中では 3 種類) に分けられている．このような多モードネットワークの例として，共通の趣味を持つ

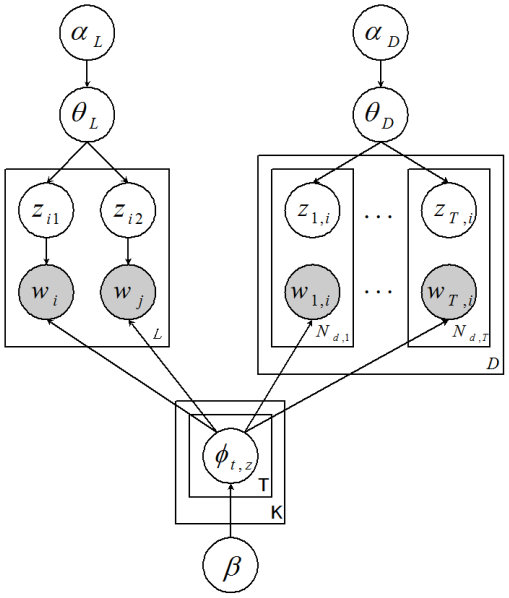

図 4　Grephical model representation of Block-LDA.

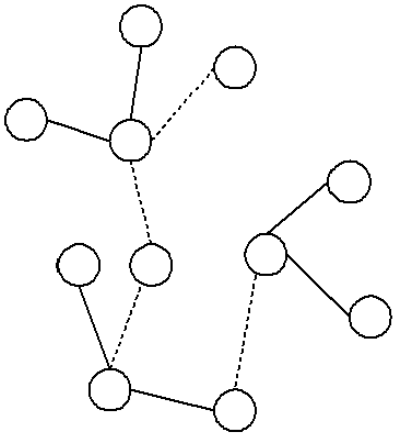

図 5　2-mode network.

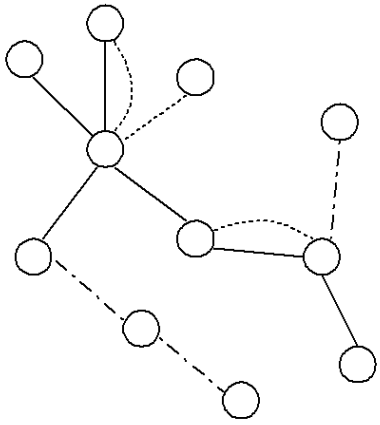

図 6　multi-mode network.

人間関係や共通の出身校である人間関係などからなるソーシャルネットワークや検索クエリログがある．

しかし，2.3 節のスパースブロックモデルや 2.4 節のブロックトピックモデルは通常のネットワークのような均質なリンクに対して提案されており，そのまま適用するのは困難である．この研究では多モードネットワークに適用しこの問題を解決するために，スパースブロックモデルを拡張した多モード・ブロックモデルを提案する．2.2 節の BTM ではなくスパースブロックモデルを拡張したのは，リンクの両端の 2 つのエンティティが常に同じトピックに属するとは実際には考えにくいからである．このモデルは多モードネットワークの各辺が T 種類に分けられるとき，各種類のリンクが共通のトピックの分布から生成されると考える．図 7 は $T = 3$ のときのトライモード・ブロックモデルのグラフィカルモデルを示している．トライモード・ブロックモデルの生成過程を以下に示す:

（1） タイプ t のリンク全体に対し，$\theta_t \sim Dirichlet(\alpha_t)$ を選択する．

（2） K 個のトピックに対し，$\phi_k \sim Dirichlet(\beta)$ を選択する．

（3） タイプ t のリンク $\ell_t = (i, j)$ に対し:

a　トピックペア $z_{\ell_t} \sim Multinomial(\theta_t)$ を選択する．

b　バイタームの 2 単語 $w_{\ell_t i}, w_{\ell_t j}$ に対し，単語 $w_{\ell i} \sim Multinomial(\phi_{z_1}), w_{\ell j} \sim Multinomial(\phi_{z_2})$ をそれぞれ選択する．

周辺化ギブスサンプリングは 3 種類のリンクを順に推定することで行う．それぞれの種類は独立なトピックペア分布から生成されるので完全条件付き確率は以下の式で与えられる:

$$p(z_{\ell_t} \mid \mathbf{z}^{\neg \ell_t}, \mathbf{L}^{\neg \ell_t}, \boldsymbol{\alpha}, \beta) \propto (L_k^{\neg \ell_t} + \alpha_t) \cdot \frac{(N_{k_1 i}^{\neg \ell_t} + \beta)(N_{k_2 j}^{\neg \ell_t} + \beta)}{(N_{k_1 \cdot}^{\neg \ell_t} + N_t \beta)(N_{k_2 \cdot}^{\neg \ell_t} + N_t \beta + \delta_k)} \tag{6}$$

ここで $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_T\}$ であり，「$\neg \ell_t$」はタイプ t のリンク ℓ を除くという事を示し，L_{ki} は単語 i がトピック k に割り当てられた回数，$N_{k\cdot}$ はトピック k が割り当てられた回数，N_t はタイプ t のリンク数を示す．また，$\mathbf{L}$ はリンクの集合を表し，図中の $|\mathbf{L}|$ はその総数を表す．

3.2　検索クエリログへの適用

ウェブ検索クエリログに対して多モード・ブロックモデルの適用を行う．検索クエリログの中で，共通のウェブページに対してクリックが行われたクエリの対や，同一の検索セッションにおいて投入されたクエリの対，部分的に共通した文字列からなるクエリの対などはそれぞれ関係があるクエリと言う事ができ，そのようなクエリ間の複雑な関係を多次元関係データあるいは多モードネットワークと見なすことができる．

4 章では現実の検索クエリログのデータを用いた実験を行い，多モード・ブロックモデルの有効性を検証する．

4.　実　　験

4.1　データセット

この研究で用いるデータセットは「Yahoo!　検索」の 3 種類の検索関連クエリデータ (注1) である．それぞれ共クリッククエリ，共トピッククエリ，共クリッククエリと呼ばれ，共ト

(注1)：http://research.nii.ac.jp/ntcir/news-20150717-ja.html

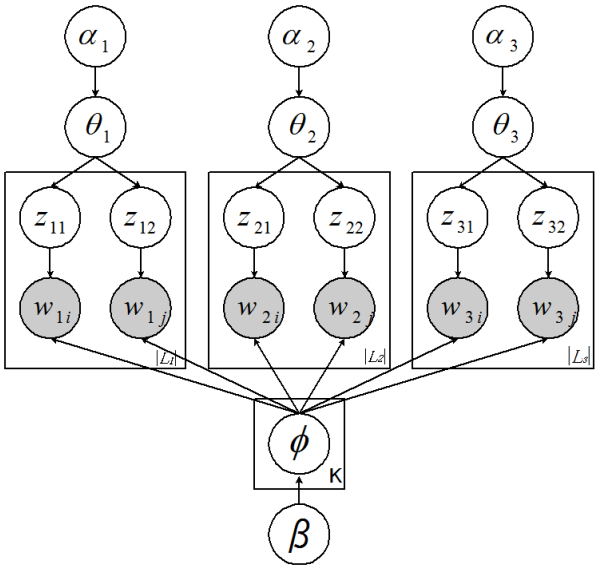

図 7　Grephical model representation of Tri-mode block model.

表 1　Statistics of the dataset used.

Query Type	Number of rerated queries
Co-Click Query	83,928
Co-Topic Query	88,075
Co-Session Query	48,768

ピッククエリと共セッションクエリを 2009 年 7 月から 2013 年 6 月，共クリッククエリは 2009 年 7 月から 2010 年 12 月の Yahoo!JAPAN 検索から抽出されている．それぞれのデータにはクエリ，関連クエリ，関連性の強さの値が含まれる．発生回数の少ないクエリはプライバシーの問題のためデータから削除されており，そのカットオフ閾値は開示されていない．各クエリについて最大 10,000 の関連するクエリ記録が各種類について含まれている．各クエリの正確な発生回数は明らかでないので，各データセットで定義される共起確率 P_{CC}, P_{CT}, P_{CS} を 1000 倍し，小数第一位を四捨五入した回数共起 (検索) されたと仮定した．また，共起回数が 0 回となるものについては除外し抽出した．クエリ，関連クエリには複数語や文からなるものが多く含まれていることから，スパース性を緩和する為に関連クエリを形態素解析し，意味のある単語ごとに分割し，それぞれがクエリと関連するように分解を行った．前処理後のデータセットの統計を表 1 に示す．

また，その事前実験として共クリッククエリの訓練データセットををスパースブロックモデルに用いて対称ディレクレハイパーパラメータ α とトピック数 K の推定を行った．既存の研究などから対称ディレクレハイパーパラメータ $\beta = 0.01$ とした．周辺化ギブスサンプリングの収束条件はヘルドアウト対数尤度の変化割合が負になった時とした．対称ディレクレハイパーパラメータ α は $0.1, 0.2, 0.3, 0.4, 0.5$ の 5 種類，トピック数 K は $20, 40, 60, 80, 100$ の 5 種類で組み合わせ，訓練セット 4 分割し，その 1 つを検証用セットとして周辺化ギブスサンプリングを行う交差検定を行った．表 2 はその結果をまとめたもの

表 2　Cross-validation results.

Number of topics K	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
20	-15.35	-15.46	-15.37	-15.46	-15.48
40	-15.02	-15.10	-15.10	-15.18	-15.20
60	-14.86	-14.98	-15.03	-15.14	-15.16
80	-14.78	-14.92	-15.01	-15.10	-15.19
100	-14.73	-14.88	-15.04	-15.16	-15.26

のであり，ハイパーパラメータは $\alpha = 0.1$ の時最も良い結果となり，トピック数は $K = 100$ の時最も良い結果となった．トピック数が大きくなるにつれて良い結果となったが $K = 80$ と $K = 100$ で大きな改善が見られなかったのでこれ以上大きいトピック数でも大きな改善は見込めない事から $\alpha = 0.1, K = 100$ を以降の実験で用いることとした．

4.1.1　共クリッククエリ

クエリ q の URL 集合 UC_q をクエリ q に対してクリックされた URL のセット，URLu のクエリ集合 QC_u を URLu がクリックされたクエリのセットと定義する．共クリッククエリセット CCQ_q は UC_q の URLu が既にクリックされたことを意味し，以下のように定義される:

$$CCQ_q \equiv \cup_{u \in UC_q} QC_q \tag{7}$$

クエリとそれに関連するクエリの間の関連性の強さを以下の重み付け方式に従って推定する．$cnt(u;q)$ をクエリ q に応じた u のクリック数，$cnt(q)$ をクエリ q に応じたクリック総数，$cnt(u)$ をクエリに関係なく u のクリック総数，Q を全てのクエリのセットと定義する．確率 $P_{CC}(q_2 \mid q_1)$ を以下のように定義する:

$$\begin{aligned} P_{CC}(q_2 \mid q_1) &= \sum_{u \in UC_{q_1}} P(u \mid q_1) \cdot P(q_2 \mid u) \\ &= \sum_{u \in UC_{q_1}} P(u \mid q_1) \cdot \frac{P(q_2) \cdot P(u \mid q_2)}{P(u)} \end{aligned}$$

4.1.2　共トピッククエリ

共トピッククエリをある項で加えられることで拡大したクエリのセットとして定義すると以下のようになる:

$$CTQ_q \equiv \alpha q \beta \mid \alpha, \beta \in term^*$$

ここで，$term^*$ は NULL 配列を含む項の配列を示す．共トピックの関係について，以下の確率でオリジナルクエリ q_1 と共トピッククエリ q_2 間の関係の強さを切り取る重み付け方式を定義する:

$$P_{CT}(q_2 \mid q_1) = \frac{cnt(q_2)}{cnt(q_1) + \sum_{q_{2'} \in CTQ_{q_1}} cnt(q_{2'})}$$

ここで，$cnt(*)$ はログ内のクエリ $*$ の発生回数を意味する．

4.1.3　共セッションクエリ

CSQ_{q_1} のセットを q_1 とセッションを共有するクエリのセットとして定義する．つまり，読み手は協調フィルタリングの文脈ではこのような関係に詳しい:また，ユーザはこれらのクエリ

を使用して何かを検索し，このクエリを使用して他の何かを検索する．

$$CSQ_{q_1} \equiv \{q_2 \mid \exists u : submit(q_1, t_1, u) \\ \wedge\ submit(q_2; q_1, t_2, u) \\ \wedge\ 0 < (t_2 - t_1) < 300s\}$$

ここで $submit(q, t, u)$ はユーザ t が時間 t にクエリ q を検索したことを示し，$submit(q_2; q_1, t_2, u)$ は q_1 の検索のすぐ後に続いて，ユーザ t が時間 t_2 にクエリ q_2 を検索したことを示している．同じユーザの検索クエリ配列中のセッションの境界を検出する代わりに，セッションを 5 分間の枠として考える．q_1 と q_2 の間の共セッション関係の強さは確率として推定される:

$$P_{CS}(q_2 \mid q_1) = \frac{cnt_{csr}(q_2; q_1)}{cnt(q_1)}$$

ここで $cnt_{csr}(q_2; q_1)$ は同じユーザのセッション内でクエリ q_2 が直接クエリ q_1 に続く数 (共セッション関係) を示す．

4.2 ハイパーパラメータ等の設定

3 種類の検索クエリデータセットをそれぞれクエリレベルで 80%訓練セットと 20%テストセットにランダムに分割し，テストセットは 3 つをまとめた 1 つのものとして扱うものとした．我々は最初に訓練セットにおいて，周辺化ギブスサンプリングを用いて共クリッククエリのみを用いるスパースブロックモデルと，3 種類の検索クエリを用いるトライブロックモデル，そしてベースラインとしてスパースブロックモデルに 3 種類の検索クエリを区別せず 1 つにまとめて用いたもの (Sparse block model-3) を推定した．

4.3 関連クエリ予測

各モデルを用いてクエリが与えられたときに関連クエリを候補の中から予測する．候補はテストセットの全ての関連クエリとし，クエリ毎に尤度のランキングを作成し，その Mean Average Precision(MAP) を計算する．MAP とそのサンプル標準偏差は以下の表 3 で示される．ベースラインとなる Sparse block model-3 と比べると提案手法であるトライブロックモデルの方が優れた結果となった．

次に 3 種類の検索クエリデータセットのうち最もユーザの検索意図が同じものが共起している確率の高いものは共クリッククエリであると仮定し，共クリッククエリの 20 % をテストセットとした場合の関連クエリ予測を行った．MAP とそのサンプル標準偏差は以下の表 4 で示される．ベースラインの Sparse block model-3 や比較手法である Sparse block model と比べると提案手法である多モード・ブロックモデルの方が優れた結果となった．比較手法である Sparse block model は共クリッククエリのみを訓練データセットとして用いて予測を行っている事から，提案手法は他の 2 種類のクエリの関連データを用いて予測精度を向上させられていると考えられる．

表 3 Mean average precision(MAP) results for general query term prediction.

	MAP	Sample Standard Deviation
Sparse block model-3	0.0362	0.0064
Tri-mode block model	0.0430	0.0049

表 4 Mean average precision(MAP) resultsfor co-click query term prediction.

	MAP	Sample Standard Deviation
Sparse block model	0.0448	0.0669
Sparse block model-3	0.0263	0.0389
Tri-mode block model	0.0542	0.0070

5. おわりに

本論文では，スパースブロックモデルを拡張することで，多モードネットワークに利用可能なトライブロックモデルを提案し，その性能を比較した．実験を通してトライブロックモデルの拡張性を示した．

本研究における今後の課題としては，予測精度を向上させるため，双対分解 [7] を用いて多目的最適化を行い，より高度な推定を行う事が考えられる．また，今回の研究では多モードネットワークの例として検索クエリのデータを実験に用いたが，他の多モードネットワークデータに対して評価を行い有用性を確認する必要がある．

謝　辞

本研究の一部は科学研究費補助金基盤研究 (B) (15H02703) の援助による．

文　献

[1] Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, Tom Mitchell "Text Classification from Labeled and Unlabeled Documents using EM", *Machine Learning - Special issue on information retrieval*,Vol.39,pp.103-134(2000)

[2] Papadimitriou Christos, Raghavan Prabhakar, Tamaki Hisao, Vempala Santosh, "Latent Semantic Indexing: A probabilistic analysis", *PODS '98 Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*,pp.159-168(1998)

[3] Blei,D.M.,Ng,A.Y., "Latent Dirichlet allocation", *Journal of Machine Learning Research*,Vol.3,pp.993-1022(2003)

[4] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng, "A Biterm Topic Model for Short Texts", *WWW '13 Proceedings of the 22nd international conference on World Wide Web*,pp.1445-1456(2013)

[5] Juuso Parkkinen, Adam Gyenge,Janne Sinkkonen, Samuel Kaski, "A block model suitable for sparse graphs", *Proceedings of the 7th International Workshop on Mining and Learning with Graphs*,(2009)

[6] Ramnath Balasubramanyan, William W. Cohen, "Block-LDA: Jointly modeling entity-annotated text and entity-entity links", *Chapman and Hall/CRC 2014*,Vol.3,pp.255-273(2014)

[7] Ravindra K. Ahuja, Thomas L. Magnanti, James B. Orlin, "Network Flows: Theory, Algorithms, and Applications", *citeulike.org*,(1993)