情報処理学会論文誌　データベース　Vol.10　No.4　21–25 (Dec. 2017)

テクニカルノート

多次元関係モデルによる利用者の情報行動に関する予測

梅原 頌平[1,a)]　江口 浩二[1,b)]

受付日 2017年6月9日, 採録日 2017年8月2日

概要：多次元関係データは多モード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，辺はそれらの関係に対応づけられ，辺には複数種類の属性のうちの 1 つが付与される．利用者の検索行動が記録されたウェブ検索クエリログについても上記のような多モードネットワークで表すことができ，頂点はクエリ，属性付きの辺はクエリ間の関係に対応づけられる．このとき，辺の属性はいくつかの仮定に基づき，たとえば 2 つの異なるクエリから得られた検索結果リストからユーザが同じウェブページを選択した場合にこれら 2 つのクエリは互いに関連すると仮定する．以上に述べた複雑なデータを解析するため，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．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，検索クエリログ，多モードネットワーク，潜在変数モデル

Predicting Users' Search Behavior Using Stochastic Multi-mode Network Models

Shohei Umehara[1,a)]　Koji Eguchi[1,b)]

Received: June 9, 2017, Accepted: August 2, 2017

Abstract: Multidimensional relationships can be represented as a multi-mode network or graph, where each node or vertex corresponds to an object, and each link or edge is attributed to one of the multiple types of relationship between a pair of objects. Web search log includes users' search behavior and can also be represented as such a multi-mode network, where each node corresponds to a query and each attributed link corresponds to a relationship between queries. The relational attributes can be derived from multiple assumptions, for instance, two queries are considered to be related to each other when two different users input those queries and click through from respective search result lists to the same Web pages. In order to analyze such complex data, this paper proposes a new multi-mode block model based on latent variable modeling. We evaluate the effectiveness of our multi-mode block model through experiments with real search query log.

Keywords: search behavior analysis, query suggestions, Web search log, multi-mode networks, and latent variable models

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．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．

[1] 神戸大学大学院システム情報学研究科
Graduate School of System Informatics, Kobe University, Kobe, Hyogo 657–8501, Japan
[a)] umehara@cs25.scitec.kobe-u.ac.jp
[b)] eguchi@port.kobe-u.ac.jp

さて，種々の離散データ集合の分析手段の 1 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．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] があげられる．

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 1 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] や，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．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とらえることはできない．

以上の問題を解決するために，本論文では新たに多モード・ブロックモデル（multi-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，多モード・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およびバイタームトピックモデル，スパースブロ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 mixture models）は，文書ごと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 の延べ単語数，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 に関する単語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．

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とらえるものである

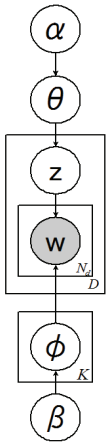

図 1　混合多項分布のグラフィカルモデル
Fig. 1　Grephical model representation of multinomial mixture model.

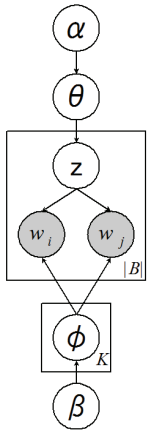

図 2　BTM のグラフィカルモデル
Fig. 2　Grephical model representation of BTM.

ため，短いテキストでは単語共起性が疎になるという問題をかかえている．そこで BTM はコーパス全体の単語共起性をとら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オブジェクト対（ここでは同じテキストデータに出現する単語対）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．**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．

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．**L** はリンクの集合を表し，図中の $|\mathbf{L}|$ はその総数を

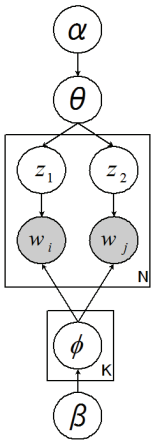

図 3　スパースブロックモデルのグラフィカルモデル
Fig. 3　Grephical model representation of Sparse Block Model.

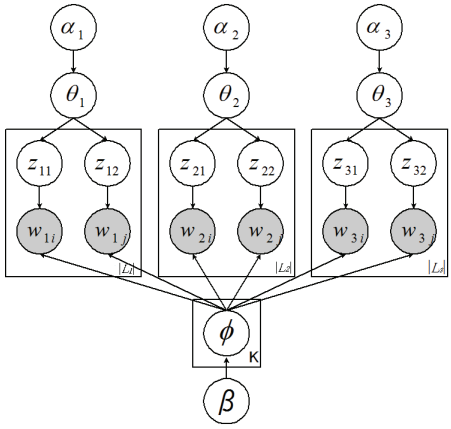

図 4　モード数を 3 としたときの多モード・ブロックモデル
Fig. 4　Multi-mode block model when the number of modes is three.

表す．

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 2 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．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．本論文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．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対の分布から生成されると考える．**図 4** は $T = 3$ のときの多モード・ブロックモデルのグラフィカルモデルを示している．多モード・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　トピック対 $z_{\ell_t} \sim Multinomial(\theta_t)$ 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対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．

$$p(z_{\ell_t} \mid \mathbf{z}^{\neg \ell_t}, \mathbf{L}^{\neg \ell_t}, \boldsymbol{\alpha}, \beta) \propto (L_k^{\neg \ell_t} + \alpha_t) \cdot \frac{(N_{k_1 i}^{\neg \ell_t} + \beta)(N_{k_2 j}^{\neg \ell_t} + \beta)}{(N_{k_1 \cdot}^{\neg \ell_t} + N_t \beta)(N_{k_2 \cdot}^{\neg \ell_t} + N_t \beta + \delta_k)} \quad (1)$$

ここで $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_T\}$ であり，「$\neg \ell_t$」はタイプ t のリンク ℓ を除くということを示し，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いうことができ，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[*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，

[*1] http://research.nii.ac.jp/ntcir/news-20150717-ja.html

表 1　データセット数
Table 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

共トピッククエリと共セッションクエリは 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．発生頻度の少ないクエリはプライバシ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．各クエリ対について最大 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頻度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 1,000 倍し，小数第 1 位を四捨五入した頻度だけ共起したと仮定した．また，共起頻度が 0 回となるものについては除外した．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ため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．

4.2　ハイパーパラメータなどの設定

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 種類のテストセットを 1 つにをまとめたものとした．

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多モード・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．

また，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 $\alpha = 0.1$，$\beta = 0.01$，$K = 100$ とした．

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すべての関連クエリとし，クエリごと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を**表 2** に示す．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多モード・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を**表 3** に示す．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ことから，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．また，2 つの実験結果についてそれぞれ Paird-T 検定を行った結果，有意水準 1%ですべての結果に有意差があることが認められた．

表 2　関連クエリ予測の MAP の結果
Table 2　Mean average precision (MAP) results for general query term prediction.

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

表 3　共クリッククエリの関連クエリ予測の MAP の結果
Table 3　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
Multi-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多モード・ブロックモデルを提案し，その性能を比較した．実験を通して多モード・ブロックモデルの有効性を示した．

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 [6] を用いて多目的最適化を行い，より高度な推定を行うことが考えられる．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．

参考文献

[1] Nigam, K., Mccallum, A.K., Thrun, S. and Mitchell, T.: Text Classification from Labeled and Unlabeled Documents using EM, *Machine Learning – Special issue on information retrieval*, Vol.39, pp.103–134 (2000).
[2] Christos, P., Prabhakar, R., Tamaki, H. and Santosh, V.: Latent Semantic Indexing: A probabilistic analy-

sis, *PODS '98 Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pp.159–168 (1998).
[3] Blei, D.M. and Ng, A.Y.: Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
[4] Yan, X., Guo, J., Lan, Y. and Cheng, X.: A Biterm Topic Model for Short Texts, *WWW '13 Proceedings of the 22nd international conference on World Wide Web*, pp.1445–1456 (2013).
[5] Parkkinen, J., Gyenge, A., Sinkkonen, J. and Kaski, S.: A block model suitable for sparse graphs, *Proc. 7th International Workshop on Mining and Learning with Graphs* (2009).
[6] Ahuja, R.K., Magnanti, T.L. and Orlin, J.B.: *Network Flows: Theory, Algorithms, and Applications*, citeulike.org (1993).

梅原 頌平

平成 28 年神戸大学工学部情報知能工学科卒業．同年より，同大学大学院システム情報学研究科情報科学専攻博士前期課程に在学中．

江口 浩二

神戸大学大学院システム情報学研究科准教授．博士（工学）．情報検索，統計的機械学習，データマイニングの研究に従事．

（担当編集委員　張 建偉）