様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　１４　日現在

機関番号：　１４５０１
研究種目：　基盤研究(B)
研究期間：　2011～2014
課題番号：　２３３０００３９

研究課題名（和文）大規模構造データに対する確率モデル推定に基づく知識の創出と活用

研究課題名（英文）Knowledge generation and use based on probabilistic model estimation from large-scale structured data

研究代表者

江口　浩二（EGUCHI, Koji）

神戸大学・システム情報学研究科・准教授

研究者番号：５０３２１５７６

交付決定額（研究期間全体）：（直接経費）　　14,800,000円

研究成果の概要（和文）：本課題では、内部構造や外部構造を持つテキストデータとネットワークデータに対して確率的に表現された潜在構造を推定する技術を開発する。ここでいう内部構造とは、たとえば、テキストデータにおいてトークン（単語）が属性で特徴づけられたものを指し、ネットワークデータにおいては各頂点または辺が属性で特徴づけられたものを指す。また、外部構造とは、たとえば、所与のネットワーク構造における各頂点にテキストデータ群が関連付けられた状況を指す。このような複雑な構造をもつ大規模なデータから低次元の潜在構造を推定することで、様々な実問題に利用可能な「知識」を抽出する。情報の検索、推薦、予測と、時系列解析などに応用する。

研究成果の概要（英文）：We aim to estimate probabilistic latent structure underlying collections of text data and network data with internal structure or external structure. The internal structure indicates, for instance, the attributed word tokens in text data or attributed nodes or edges in network data. The external structure indicates, for instance, the case when each node of network data is associated with a set of text data. We extract "knowledge" that can apply for various real-world problems, by estimating low-dimensional latent structure from a large amount of complexly structured data. We apply our techniques to the problems of information retrieval, recommendation, prediction, and time-series analysis.

研究分野：　情報学

キーワード：　統計モデリング　混合メンバシップモデル　トピックモデル　統計的ネットワークモデル　潜在変数モデル　ギブスサンプリング　パーティクルフィルタ

１．研究開始当初の背景

テキストやネットワーク（グラフ）などの離散データの分析の手段として、潜在トピックモデルや混合メンバシップモデルなどと呼ばれるアプローチが注目されつつある。とりわけ、テキストデータに対しては潜在的ディリクレ配分法(LDA: Latent Dirichlet Allocation)、ネットワークデータに対しては混合メンバシップ・ブロックモデル（MMSB: Mixed Membership Stochastic Blockmodels)、また、潜在トピック数をデータから推定する仕組みを備えた階層ディリクレ過程（HDP: Hierarchical Dirichlet processes）が典型である。

ところが、以上のようなモデルを現実の複雑な構造をもつデータに適用するには自明でない拡張が不可欠である。また、現実の問題への応用や、大規模なデータへの対処など、重要な課題が少なくなかった。

２．研究の目的

本課題では、内部構造や外部構造を持つテキストデータとネットワークデータに対して確率的な潜在構造を推定する技術を開発する。ここでいう内部構造とは、たとえば、テキストデータにおいてトークン（単語）が属性で特徴づけられたものを指し、ネットワークデータにおいては各頂点または辺が属性で特徴づけられたものを指す。また、外部構造とは、たとえば、所与のネットワーク構造における各頂点にテキストデータ群が関連付けられた状況を指す。本課題は、以上に述べたような複雑な構造をもつ大規模なデータから低次元の潜在構造を推定することで、様々な実問題の解決に利用可能な「知識」を抽出することを目指すものである。具体的な応用問題として、情報の検索、推薦、予測と、時系列解析に着目する。

３．研究の方法

(1) 属性付きテキストデータとして、Wikipedia を典型として言語間比較可能データを取り上げ、各言語をモードと見なしたマルチモーダルなデータとして一般化する。それに対して、LDA に基づくトピックモデルを拡張したマルチモーダルトピックモデルを開発する。また、映像データなどの他のマルチモーダルデータについても検討する。

(2) 頂点または辺に属性が付与されたネットワークデータに対して、MMSB に新たな確率変数を追加することにより、モデル化し、それらを周辺化ギブスサンプラにより推定する。応用として、情報推薦問題などに適用する。また、ネットワークの時間変化を考慮した推定を実現するため、パーティクルフィルタに基づく新たな推定手法を開発する。

(3) 潜在トピックに基づいてテキストデータ間の関係をモデル化する関係トピックモデルの考え方に基づいて、マルチモーダルデータにおけるモード間の関係とデータ間の関係をモデル化する。これにより、個々のマルチモーダルデータにおけるモード間の関係による内部構造表現と、データ間の関係による外部構造表現を同時にモデル化することが可能となる。また、大規模データに対する LDA や HDP の高速推定のため、分散並列環境を想定したアルゴリズムを開発する。

４．研究成果

(1) マルチモーダルトピックモデル：

① 複数の表現からなるマルチモーダルデータに着目し、モード間の依存性を的確に捉えつつ次元削減を実現する Symmetric Correspondence Topic Models なる潜在トピックモデルを提案した。本提案モデルの主な特徴の一つは、各データを構成する複数のモードから主軸となるモードを予め特定する必要性がない点にあり、その適用範囲は広い。言語間比較可能テキストデータから言語を横断して潜在トピックを推定する問題に提案モデルを適用し、その有効性を示した。

② また、映像データを構成するキーフレーム画像における視覚単語や発話単語に着目し、それらを統合する潜在トピックモデルを提案した。特に、各トピックに関して同一映像中では同じ視覚単語や発話単語による表現が頻出する傾向を考慮した。インターネット映像データのジャンル分類に関する評価実験により、提案モデルの有効性を示した。

(2) 統計的ネットワークモデル：

① 複数種類のノードまたはリンクで構成されたネットワークの潜在構造をモデル化した。また、それを用いて、ユーザのアイテム利用履歴に基づいて構成された二部グラフと社会的ネットワークなどを合成して得られた異種ネットワークに適用し、上位 N 推薦問題における提案モデルの有効性を示した。

② また、大規模で時間的変化を伴うネットワーク表現の潜在構造を推定し、未知の関係を予測することを目的として、MMSB を高速かつ高精度に推定する忘却型パーティクルフィルタを提案した。提案手法は、ネットワークにおいて時間経過にともなって頂点やリンクの追加が頻繁に起こり得ることと、ネットワークの潜在的な構造に関して時間変化が起こり得ることを考慮している。社会ネットワ

ークデータを用いた評価実験により、提案手法の有効性を示す結果を得た。

(3) 複雑な構造を持つデータ、大規模なデータの分析：

① 潜在トピックに基づいてマルチモーダルデータにおけるモード間の関係およびデータ間の関係をモデル化するマルチモーダル関係トピックモデルを開発し、関係予測性能ならびにテストデータ予測性能に関する評価実験を行った。

② 潜在変数を用いたデータ解析手法である LDA および HDP に関して、大規模データのために高速な推定を実現する分散並列アルゴリズムを開発し、クラウド環境などを用いた評価実験を行った。その結果、提案した推定手法により、予測性能を維持しつつ高い効率性が実現することを明らかにした。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 17 件）

[1] Quang-Hong Vuong, Atsuhiro Takasu: "Transfer Learning for Emotional Polarity Classification", Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, Vol.2, pp.94-101 (2014). 査読有.
DOI: 10.1109/WI-IAT.2014.85

[2] Yang Xie, Koji Eguchi: "Multimedia Topic Models Considering Burstiness of Local Features", IEICE Transactions on Information and Systems, Vol.E97-D, No. 4, pp.714-720 (2014). 査読有.
DOI: 10.1587/transinf.E97.D.714

[3] Tomoki Kobayashi, Koji Eguchi: "Online Inference of Mixed Membership Stochastic Blockmodels for Network Data Streams", IEICE Transactions on Information and Systems, Vol.E97-D, No.4, pp.752-761 (2014). 査読有.
DOI: 10.1587/transinf.E97.D.752

[4] Tsukasa Omoto, Koji Eguchi, Shotaro Tora: "Hybrid Parallel Inference for Hierarchical Dirichlet Processes", IEICE Transactions on Information and Systems, Vol.E97-D, No.4, pp.815-820 (2014). 査読有.
DOI: 10.1587/transinf.E97.D.815

[5] Atsuhiro Takasu, Manabu Ohta: "Rule Management for Information Extraction from Title Pages of Academic Papers", Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods, pp.438-444 (2014), 査読有.
DOI: 10.5220/0004827204380444

[6] Tomoki Kobayashi, Koji Eguchi: "Sequential Monte Carlo Inference of Mixed Membership Stochastic Blockmodels for Dynamic Social Networks", NIPS 2013 Workshop on Frontiers of Network Analysis: Methods, Models, and Applications (2013). 査読有.
http://arxiv.org/abs/1312.2154

[7] Pannawit Samatthiyadikun, Atsuhiro Takasu, Saranya Maneeroj: "Bayesian Model for a Multicriteria Recommender System with Support Vector Regression", Proceedings of the 14th IEEE International Conference on Information Reuse and Integration, pp.38-45 (2013), 査読有.
DOI: 10.1109/IRI.2013.6642451

[8] 内藤 慎也, 江口 浩二: "閲覧履歴グラフに基づく正則化リンク解析を用いたロバスト推薦", 日本データベース学会論文誌, Vol.12, No.1, pp.7-12 (2013). 査読有.
http://dbsj.org/journal/dbsj_journal/dbsj_journal_vol_12_no_1_7_12/

[9] 石黒 七海, 江口 浩二, 横峯 樹: "異種混合メンバーシップ・ブロックモデルと情報推薦への応用", 日本データベース学会論文誌, Vol.12, No.1, pp.43-48 (2013). 査読有.
http://dbsj.org/journal/dbsj_journal/dbsj_journal_vol_12_no_1_43_48/

[10] Tsukasa Omoto, Koji Eguchi, Shotaro Tora: "Hybrid Parallel Inference for Hierarchical Dirichlet Process", ICML 2013 workshop on Inferning: Interactions between Inference and Learning (2013). 査読有.
http://openreview.net/file/14550ac7-2739-4f2d-b261-ab68492db8dd.pdf

[11] Shotaro Tora, Koji Eguchi: "MPI/OpenMP Hybrid Parallel Inference Methods for Latent Dirichlet Allocation: Approximation and Evaluation", IEICE Transactions on Information and Systems, Vol.E96-D,

No.5, pp.1006-1015 (2013). 査読有.
DOI: 10.1587/transinf.E96.D.1006

[12] Kosuke Fukumasu, Koji Eguchi, Eric P. Xing: "Symmetric Correspondence Topic Models for Multilingual Text Analysis", Advances in Neural Information Processing Systems 25 (NIPS 2012), pp.1295-1303 (2012). 査読有.
http://papers.nips.cc/paper/4583-symmetric-correspondence-topic-models-for-multilingual-text-analysis

[13] Pannawit Samatthiyadikun, Atsuhiro Takasu, Saranya Maneeroj: "Multicriteria Collaborative Filtering by Bayesian Model-based User Profiling", Proceedings of the 13th IEEE International Conference on Information Reuse and Integration, pp.124-131 (2012), 査読有.
DOI: 10.1109/IRI.2012.6303000

[14] Shinya Naito, Koji Eguchi: "Robust Recommendations using Regularized Link Analysis of Browsing Behavior Graphs", Social Computing, Behavioral-Cultural Modeling and Prediction: 5th International Conference, SBP 2012, College Park, MD, USA, Lecture Notes in Computer Science, Vol.7227, pp.339-347, Springer (2012). 査読有.
DOI: 10.1007/978-3-642-29047-3_41

[15] Atsuhiro Takasu: "A Multicriteria Recommendation Method from Data with Missing Rating Scores", Proceedings of the 2011 International Conference on Data and Knowledge Engineering, pp.60-67 (2011), 査読有.
DOI: 10.1109/ICDKE.2011.6053931

[16] Shotaro Tora, Koji Eguchi: "MPI/OpenMP Hybrid Parallel Inference for Latent Dirichlet Allocation", Proceedings of the 3rd Workshop on Large Scale Data Mining: Theory and Applications, pp.1-7 (2011). 査読有.
DOI: 10.1145/2002945.2002950

[17] Atsuhiro Takasu, Saranya Maneeroj: "A Recommendation Algorithm Using Positive and Negative Latent Models", Proceedings of the 2011 IEEE Symposium on Computational Intelligence and Data Mining, pp.72-79 (2011), 査読有.
DOI: 10.1109/CIDM.2011.5949455

〔学会発表〕(計 22 件)

[1] Koji Eguchi: "Multi-modal Topic Models and Large-scale Data Analysis" (Invited Talk), The 9th Korea-Japan Database Workshop 2014, 2014 年 11 月 29 日-2014 年 12 月 2 日, Gangwon (大韓民国).

[2] 上野 良輔, 江口 浩二: "双対分解を用いたマルチタスク最大マージントピックモデル", 第 16 回情報論的学習理論ワークショップ, 2014 年 11 月 16 日-2014 年 11 月 19 日, 名古屋大学(愛知県).

[3] 西出 飛翔, 江口 浩二: "無限潜在特徴関係モデルのマージン最大化推定による離散関係属性付きネットワークの分析", 2014 年度情報処理学会関西支部大会, No.C-10, 2014 年 9 月 17 日, 大阪大学中之島センター(大阪府).

[4] 坂田 洋介, 江口 浩二: "マージン最大化マルチモーダル関係トピックモデルと多言語間関係予測による評価", 2014 年度情報処理学会関西支部大会, No.F-01, 2014 年 9 月 17 日, 大阪大学中之島センター(大阪府).

[5] 島廻 卓史, 江口 浩二: "アノテーション付き画像の潜在トピック階層に関するノンパラメトリックベイズモデリング", 2014 年度情報処理学会関西支部大会, No.G-05, 2014 年 9 月 17 日, 大阪大学中之島センター(大阪府).

[6] 山本 浩平, 江口 浩二, 高須 淳宏: "カテゴリ階層の拡張を目的とした階層的トピックモデル", 第 6 回データ工学と情報マネジメントに関するフォーラム. 2014 年 3 月 3 日-2014 年 3 月 5 日, 淡路夢舞台(兵庫県).

[7] 坂田 洋介, 江口 浩二: "マルチモーダル関係トピックモデルによる多言語間関係予測", 第 6 回データ工学と情報マネジメントに関するフォーラム. 2014 年 3 月 3 日-2014 年 3 月 5 日, 淡路夢舞台(兵庫県).

[8] 山本 浩平, 江口 浩二, 高須 淳宏: "カテゴリ階層の拡張を目的とした半教師あり階層的トピックモデル", 第 16 回情報論的学習理論ワークショップ, 2013 年 11 月 11 日-2013 年 11 月 13 日, 東京工業大学(東京都).

[9] 小林 知己，江口 浩二："時間変化を伴うネットワークにおける混合メンバシップ・ブロックモデルのオンライン学習"，第16回情報論的学習理論ワークショップ，2013年11月11日-2013年11月13日，東京工業大学(東京都)．

[10] 大元 司，江口 浩二，東羅 翔太郎："階層ディリクレ過程のハイブリッド並列化推定"，第16回情報論的学習理論ワークショップ，2013年11月11日-2013年11月13日，東京工業大学(東京都)．

[11] 上野 良輔，江口 浩二："回帰分析のためのマージン最大化トピックモデルのギブスサンプリング推定"，電子情報通信学会技術研究報告，Vol.113，No.150，DE2013-31，pp.187-192，2013年7月22日-2013年7月23日，北海道大学(北海道)．

[12] 謝 洋，江口 浩二："映像データにおける局所特徴のバースト性を考慮したトピックモデリング"，電子情報通信学会技術研究報告，Vol.113，No.75，PRMU2013-20，pp.5-10，2013年6月10日-2013年6月11日，NTT武蔵野研究開発センタ(東京都)．

[13] 石黒 七海，江口 浩二，横峯 樹："異種混合メンバーシップ・ブロックモデルと情報推薦への応用"，第5回データ工学と情報マネジメントに関するフォーラム，2013年3月3日-2013年3月5日，ホテル華の湯(福島県)．

[14] 大元 司，東羅 翔太郎，江口 浩二："大規模データのための階層ディリクレ過程の並列推定"，第5回データ工学と情報マネジメントに関するフォーラム，2013年3月3日-2013年3月5日，ホテル華の湯(福島県)．

[15] 小林 知己，江口 浩二："混合メンバシップ・ブロックモデルのオンライン学習"，第4回データ工学と情報マネジメントに関するフォーラム，2012年3月3日-2012年3月5日，シーサイドホテル舞子ビラ神戸(兵庫県)．

[16] 山本 浩平，江口 浩二，高須 淳宏："カテゴリ階層を考慮した確率的トピックモデルのモデル選択付き学習"，第4回データ工学と情報マネジメントに関するフォーラム，2012年3月3日-2012年3月5日，シーサイドホテル舞子ビラ神戸(兵庫県)．

[17] 下田 敬祐，江口 浩二："多腕バンディットによる検索結果の多様化に関する大規模クリックスルーログを用いた評価"，第4回データ工学と情報マネジメントに関するフォーラム，2012年3月3日-2012年3月5日，シーサイドホテル舞子ビラ神戸(兵庫県)．

[18] 西原 聖志，江口 浩二："局所特徴のバースト性を考慮した異種メディアの統合的なモデル化"，第4回データ工学と情報マネジメントに関するフォーラム，2012年3月3日-2012年3月5日，シーサイドホテル舞子ビラ神戸(兵庫県)．

[19] 横峯 樹，江口 浩二："マルチタイプ混合メンバーシップ・ブロックモデルを用いた情報推薦"，第14回情報論的学習理論ワークショップ，2011年11月9日-2011年11月11日，奈良女子大学(奈良県)．

[20] 福増 康佑，松浦 愛美，江口 浩二："Symmetric Correspondence Topic Modelsによる多言語トピック抽出"，第14回情報論的学習理論ワークショップ，2011年11月9日-2011年11月11日，奈良女子大学(奈良県)．

[21] 内藤 慎也，江口 浩二："正則化付きリンク構造解析を用いたコールドスタート推薦"，情報処理学会研究報告，Vol.2011-DBS-153，No.20，pp.1-8，2011年11月3日，エステック情報ビル(東京都)．

[22] 福増 康佑，松浦 愛美，江口 浩二："多言語トピックモデルによる言語横断リンク検出"，情報処理学会研究報告，Vol.2011-SLP-201/2011-SLP-86，No.4，pp.1-7，2011年5月16日，東京大学(東京都)．

〔その他〕
ホームページ等
http://www.prmir.scitec.kobe-u.ac.jp/

６．研究組織
(1)研究代表者
江口 浩二 (EGUCHI, Koji)
神戸大学・大学院・システム情報学研究科・准教授
研究者番号：50321576

(2)研究分担者

高須 淳弘 (TAKASU, Atsuhiro)
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：90216648

大川 剛直 (OHKAWA, Takenao)
神戸大学・大学院・システム情報学研究科・教授
研究者番号：30223738