様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　２１　日現在

機関番号：　１７３０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０２５６

研究課題名（和文）タイニーデータマイニング：基底としての確率分布による大規模データの再構成

研究課題名（英文）Tiny data mining: reconstruction of large scale data with probability distributions as bases

研究代表者

正田　備也（MASADA, Tomonari）

長崎大学・工学研究科・准教授

研究者番号：６０４１３９２８

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：この研究は規模の大きなデータの要約を目指しています。主に扱うのは文字で書かれたデータ、つまりテキストデータです。ニュース記事、学術論文、小説などがこれにあたります。テキストデータも量が多くなってくると、ひとつひとつ人間が目を通すわけにいかなくなります。そこで要約を作ります。この研究が作る要約は単語リストです。例えば「試合、ヒット、ピッチャー、トレード」という単語リストを見ると、私たちはこれが野球というトピックを表していると分かります。このような単語リストを膨大なテキストデータから自動的にいくつも取り出し、文章をひとつひとつ読まなくても何が書いてあるか分かるようにするのが、この研究の目的です。

研究成果の概要（英文）：The aim of our research is to make a efficient and effective summary of a large set of documents like news articles, academic papers, novels, etc. When the number of given documents is very large, we can only read a small portion of it. As a result, we may miss the documents containing our favorite topics. Therefore, our research aims to extract word lists from the give document set as a summary. For example, one among the extracted word lists was "game, hit, pitcher, and trade," we can know that there are documents discussing baseball. In this manner, by looking at the extracted word lists, we can know what kind of topics are discussed in the given document set. Furthermore, our research also provides a clue to find which documents are closely related to which word lists. Therefore, we can also find the documents relevant to the word lists we choose. While an existing method called topic modeling is adopted in our research, we propose its new application and its new implementation.

研究分野：　データマイニング

キーワード：　トピックモデル　機械学習　ベイズ推定　データマイニング　テキストマイニング

１．研究開始当初の背景

研究を始めた段階では、トピックモデルを大規模データの要約に役立てるという、応用方面の研究内容を軸として計画を立てていた。というのも、ビッグデータ分析が研究トレンドで、データサイエンスを具体的なデータへと適用して有効性を検証することの重要性が当時さけばれていたためである。特にトピックモデルは確率的な考え方に基づくデータ分析手法として良く知られており、大規模テキストデータ分析に関する研究の中心的な道具として本研究で採用した。

しかし、トピックモデルについてはそのパラメータを推定する計算をいかに大規模データへ適応させるかが従来から問題になっており、本研究を計画した段階では、OpenMP や CUDA の利用など、実装面での工夫をおこなうことを計画していた。

しかし、研究開始後、トピックモデルを含む機械学習全般をとりまく世の中の環境が大きく変わった。人工知能ブームである。そのため、深層学習を主要な道具としてビッグデータ分析をおこなう流れが主流となっていった。そこで、以下に記述するように、研究内容を当初の予定とは異なるものへと転換させる必要性がうまれてきた。

２．研究の目的

当初の研究目的は、３つの点に分けて提示できる。まず、トピックモデルを用いることによって大規模データの良い“要約”を作ることが内容面での目的であった。人間が、ビッグデータの規模に押しつぶされることなく、自分の必要な情報へと的確に注意を集中させることができるような手段を提供することが、この研究の目的であった。その要約が持つべき特徴として、Distinctiveness、Diversity、Dynamism という３つの D を掲げていた。これは大規模データ分析の内容面に関する研究目的である。

また、データ分析の実施方法に関する目的として、実装、つまりプログラミング上の工夫による高速化によってできるだけ規模の大きなデータに対処できるようにすることを掲げていた。具体的には、OpenMP や CUDA、場合によっては MPI を使うことによる高速化である。これは、ビッグデータ分析の実行手段に関する研究目的である。

さらに、抽出されたトピックが上述の３つの D の特徴をもつ要約とみなせるかどうかを調べる手順を提供することも目的のひとつだった。これは分析結果の評価に関する研究目的である。

しかし、2014 年に研究を開始して以降、機械学習をとりまく環境が急速に変化し、深層学習の流行を無視できなくなった。ただし、以下に記述するように、結果としては当初の研究目的に寄与するような成果が得られた。特に、トピックモデルの推定方法について、深層学習から新しい考え方を導入できたことは、今後の研究にもつながる成果であったし、大規模データにトピックモデルを適応させるという当初の研究目的を果たす、予想外に良い手段が見つかったとすら言える。

３．研究の方法

上述の研究目的の３つの側面に即して、研究方法を列挙していく。

まず、研究内容の面では、トピックモデル自体の高度化により、３つの D の特徴をもつ要約としてのトピックの抽出を目指した。この点では、交通流の速度の分析に関して一定の成果を得た。しかし、研究の方向性の変更にともない、後述のように研究方法もモデルそのものではなく、モデルのための推定計算を見直す方向へ変更された。

次に、実装の面では、マルチコアや GPU を用いた推定計算の並列化によりトピック抽出の高速化を目指した。MPI の利用については、下に述べるような研究内容の変更があったため、方法としては採用しなかった。しかし、GPU の利用は、当初想定していたよりも重要性を増すことになった。

最後に、評価の面では、やはり下に述べるような研究内容の変更により、定量的な評価については通常の perplexity による評価にとどまった。ただし、定性的な評価については、Python で実装されたワードクラウドを活用し、抽出されたトピックの良し悪しを直感的に判断できるようになった。

特に、最後のワードクラウドの作成についてはプログラミング言語として Python を使っている。これまで自己の研究では C 言語を中心に使ってきたが、本研究を進める中で、徐々に Python を活用するようになった。主な理由は、やはり、Python が深層学習の分野で中心的に使われているプログラミング言語となっていることである。本研究でも、まずは評価の部分で Python を使い始め、最終的には、推定計算自体も Python で実装することを試みるようになった。このように、研究の方法という点においても、機械学習をめぐるトレンドの変化が本研究に大きな影響を与えている。

４．研究成果

(1) 交通流速度のメタデータを用いたトピックモデリング[Masada+ FDSE15]

研究を始めた当初は、すでに着手していた交通流速度データ分析の手法を、計画書に書いたとおり、メタデータ、具体的には車速を計測した時刻や、計測された地点の場所情報も利用して高度化する研究に取り組んだ。その際、モデルそのものを multifaceted SAGE [Eisenstein+ 11]などのトピックモデルを参考にしつつ高度化した。さらには、推定計算

も、変分ベイズではなくスライスサンプリングを使うことで、より正確な推定ができるよう工夫した。交通流の速度データという、非常に全体を概観しづらいデータについて、有用な要約を得るという作業は、これによりうまく遂行できたと考える。この成果は FDSE 2015 に受理された論文にまとめられている。

(2) 深層学習での Reparameterization のトピックモデルへの適用

ところが、この研究を進めている間に、深層学習の方面で提案されたあるアイデアが、トピックモデルなど、ベイズ的な立場からのデータマイニングにとって非常に重要な意味を持つことに気がついた。それは、Kingma らによって提案された変分オートエンコーダである[Kingma+ 13]。特に、そこで用いられている reparameterization というテクニックが、トピックモデルをはじめとするベイズ的確率モデル全般のための事後分布推定計算にとって、無視しがたいブレイクスルーとなっていることに気がついた。

どのような意味で無視しがたかったかと言うと、この reparameterization というテクニックを使えば、トピックモデルを、あたかもニューラルネットワークのように訓練する可能性が開かれたからである。(技術的に言えば、ベイズ的確率モデルの事後分布推定における variational lower bound の最大化を、これまでよりもシンプルに勾配法で実現できるようになったため、このような可能性が開かれたわけである。)

従来のトピックモデルの推定計算では、データ集合が全体として与えられており、その集合を繰り返しスキャンすることで、推定計算を収束させていた。その際、どちらかと言えば閉じた式によって事後分布のパラメータを更新することが好まれていた。しかし、深層学習では、入力データはストリームと見なされる。もちろん、実際の評価実験では同じデータ集合を何度もスキャンすることがおこなわれるが、最適化計算そのものの中には、たえず新しい入力データが与えられるという情況を妨げるものは何もない。この設定の中で、確率的勾配法によりひたすらパラメータが更新されていく。

そこで、本研究は方針転換をおこなった。研究の目的を、トピックモデルの推定計算でこの reparameterization のテクニックが本当に使いものになるのかどうかを検証することに変更した。そうしなければ、今後、トピックモデルを主なデータ分析の道具とするにしても、時流に乗れないからである。

この方針転換による研究成果は、ICCSA 2016 と APWeb 2016 で発表された。前者では、LDA (潜在的ディリクレ配分法) について、後者では、CTM (correlated topic model) について、それぞれ新しい変分推定法を提案している。ポイントは、reparameterization のテクニックを使ったことである。下に、ICCSA 2016 の論文から、reparameterization を使うことによって得られた勾配の式を下に示しておく。

$$\frac{\partial \hat{L}(\mathbf{\Lambda})}{\partial \tau_{\theta,dk}} = \frac{1}{2} + \frac{1}{2}\exp\left(\frac{\tau_{\theta,dk}}{2}\right)\frac{\sum_{l=1}^{L}\epsilon_{\theta,dk}^{(l)}\{(N_{dk}+\alpha) - (N_d + K\alpha)\theta_{dk}^{(l)}\}}{L}$$

$$\frac{\partial \hat{L}(\mathbf{\Lambda})}{\partial \tau_{\phi,kv}} = \frac{1}{2} + \frac{1}{2}\exp\left(\frac{\tau_{\phi,kv}}{2}\right)\frac{\sum_{l=1}^{L}\epsilon_{\phi,kv}^{(l)}\{(N_{kv}+\beta) - (N_k + V\beta)\phi_{kv}^{(l)}\}}{L}$$

APWeb 2016 の論文でも同様な勾配の式が得られている。これらの研究により、いままでと同程度の perplexity を保ちつつ、あたかもニューラルネットワークを訓練するように、勾配を使ってパラメータを更新しつつ、トピックモデルを訓練することが出来そうだという確信を持った。

なお、上で評価に関して Python の有用性を述べておいたが、実際に ICCSA 2016 の論文に掲載したワードクラウドをここでも示しておく (図 1)。これは UCI の NYT データから得たトピックの可視化である。このように、本研究では抽出されたトピックの可視化による直感的な評価にも取り組んでいる。

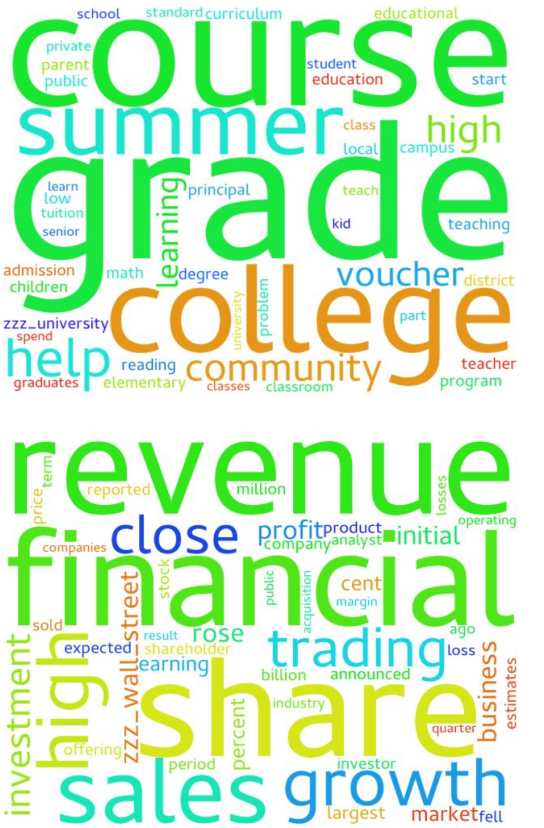

図 1. ニューヨークタイムズの記事群から得られたワードクラウド

(3) トピックモデルでの多層パーセプトロンの利用 [Masada BDM17]

本研究に残された時間で、より直接的に深層学習の成果をトピックモデルへと反映させる取り組みをおこなった。

LDA においてデータ集合全体を使って訓練されるパラメータは、トピックごとの単語確率である。一方、文書ごとのトピック確率は各文書に依存するローカルなパラメータである。上述の自己の２つの研究（ICCSA2016 と APWeb2016）でも、勾配法で更新されるパラメータは、トピックごとの単語確率、つまり、グローバルなパラメータであった。

そこで、このグローバルなパラメータを、多層パーセプトロンの出力から得ることを考えた。多層パーセプトロンは、全結合層とも呼ばれ、最も基本的なニューラルネットワークの構造である。この研究では、隠れ層がひとつという最もシンプルな多層パーセプトロンからトピックごとの単語確率分布を得ることで、LDA に関して今までと違ったことができるかどうかを確認した。その成果は、年度をまたいでしまったが、PAKDD 2017 の併設ワークショップ BDM に受理され、2017 年 5 月末に発表した。

この研究成果で特筆すべき点は、全文書数を知らなくても LDA のオンライン学習が実行できることを示したことである。提案手法では、深層学習のようにひたすらデータのストリームを撫でていきつつ、トピック毎の単語確率を更新することが可能となっている。従来研究として、[Hoffman+ 10]により提案された LDA のオンライン学習があるが、これは、その背景にある理論によって、全文書数が既知であることが必要とされていた。しかし、例えば深層学習による画像データの分析において、全画像の枚数が学習に必要という話は聞いたことがない。

本研究では、全文書数を知らなくても、collapsed Gibbs sampling による LDA の事後分布推定に近い test perplexity を達成できる場合もあることを明らかにしている。

なお、深層学習系の文書モデルとして、NVDM [Miao+ 16]というモデルもあるが、これは変分ベイズ推定の LDA には test perplexity による評価で勝てても、collapsed Gibbs sampling の LDA には勝てないことが[Srivastava+ 2017]によって示されている。つまり、やはり文書モデルとして LDA はまだ有力であり、これを改良していく意義はまだあるということである。この意味で、本研究の成果は意義があると考える。

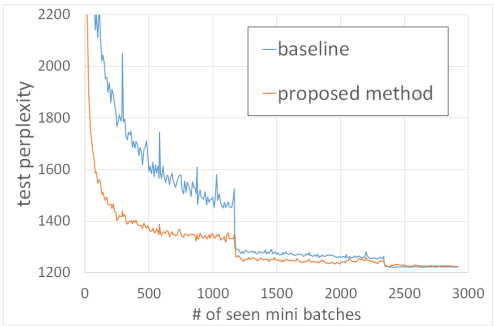

図 2．baseline と提案手法とにおける perplexity の減り方の比較

なお、図 2 に、隠れ層をひとつ含む多層パーセプトロン(proposed method)と、隠れ層なしで単にパーセプトロンから softmax で単語確率を得た場合(baseline)との、test perplexity の減り方を比較したグラフを示す。横軸は処理したミニバッチの個数である。このように、提案手法のほうが少ないミニバッチ数で大きく perplexity を改善できていることが分かる。複雑なニューラルネットを使う利点が明確に現われている。

ただし、計算時間の点では、複雑なニューラルネットのほうが不利であり、この点については、CUDA でコードを書いて GPU を利用することで部分的に克服できたが、今後は、PyTorch などの深層学習フレームワークを使うことで、より徹底して GPU の並列度を利用することを考えている。

なお、この研究についても、結果として得られたトピックをふたつ、ワードクラウドで可視化して下に示しておく。これらは、MEDLINE®データから得られたものである。

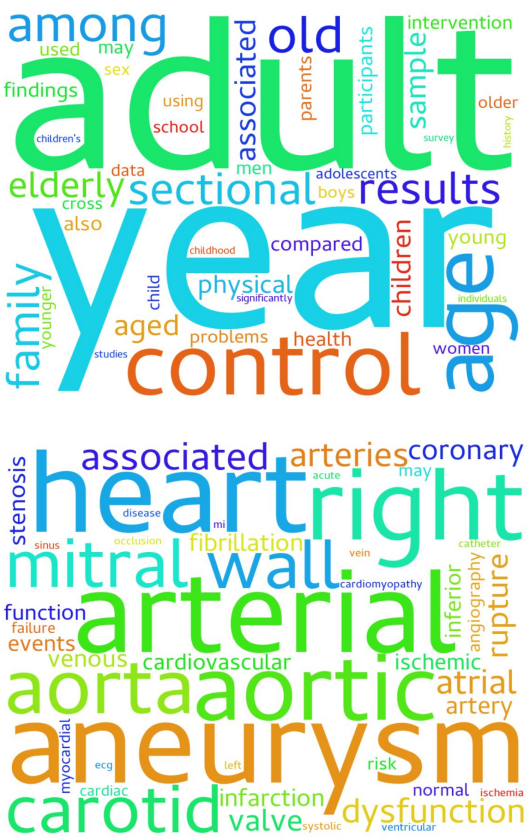

図 3．MEDLINE®データから得られたワードクラウド

＜引用文献＞

① Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *JMLR* 3, pp.993-1022, 2003.

② Hoffman, M.D., Blei, D.M., Bach, F.: Online learning for latent Dirichlet allocation. In *NIPS*, 2010.

③ Eisenstein, J., Ahmed, A., Xing, E.P.: Sparse additive generative models of text. In *ICML*, 2011.
④ Kingma, D. P., Welling, M.: Stochastic gradient VB and the variational auto-encoder. In *ICLR*, 2014.
⑤ Miao, Y., Yu, L., Blunsom, P.: Neural variational inference for text processing. In *ICML*, 2016.
⑥ Srivastava, A., Sutton, C.: Autoencoding variational inference for topic models. In *ICLR*, 2017.

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 1 件）

① Yuzana Win, Tomonari Masada: Bidirectional Extraction of Phrases for Expanding Queries in Academic Paper Retrieval. *IJARAI*, Vol.5, Issue 1, 2016 **（査読有）**
DOI: 10.14569/IJARAI.2016.050105

〔学会発表〕（計 7 件）

① Tomonari Masada: Estimating Word Probabilities with Neural Networks in Latent Dirichlet Allocation. In *PAKDD Workshops*, May 25, 2017, Jeju, Korea.
② Tomonari Masada, Atsuhiro Takasu: A Simple Stochastic Gradient Variational Bayes for the Correlated Topic Model. In *APWeb* (2) September 25, 2016: 424-428, Suzhou, China
DOI: 10.1007/978-3-319-45817-5_39
③ Tomonari Masada, Atsuhiro Takasu: A Simple Stochastic Gradient Variational Bayes for Latent Dirichlet Allocation. In *ICCSA* (4) July 6, 2016: 232-245, Beijing, China
DOI: 10.1007/978-3-319-42089-9_17
④ Yuzana Win, Tomonari Masada: Extraction of proper names from myanmar text using latent dirichlet allocation. In *TAAI*, November 25, 2016: 96-103, Hsinchu, Taiwan
DOI: 10.1109/TAAI.2016.7880176
⑤ Yuzana Win, Tomonari Masada: Exploring Technical Phrase Frames from Research Paper Titles. In *AINA Workshops*, March 27, 2015: 558-563, Gwangju, Korea
DOI: 10.1109/WAINA.2015.37
⑥ Tomonari Masada, Atsuhiro Takasu: Traffic Speed Data Investigation with Hierarchical Modeling. In *FDSE*, November 25, 2015: 123-134, Ho Chi Minh, Vietnam
DOI: 10.1007/978-3-319-26135-5_10
⑦ Tomonari Masada, Atsuhiro Takasu: Heuristic Pretraining for Topic Models. In *IEA/AIE*, June 10, 2015: 351-360, Seoul, Korea
DOI: 10.1007/978-3-319-19066-2_34

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://diversity-mining.jp/

６．研究組織
(1)研究代表者
正田　備也（MASADA, Tomonari）
長崎大学・工学研究科・准教授
研究者番号：60413928