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　　４　日現在

機関番号：　６２６１５
研究種目：　研究活動スタート支援
研究期間：　2013～2014
課題番号：　２５８８００２８

研究課題名（和文）高度な情報処理を支える大規模テンソル分解の開発

研究課題名（英文）Large-scale tensor decomposition for high-level information processing

研究代表者
林　浩平（HAYASHI, Kohei）

国立情報学研究所・ビッグデータ数理国際研究センター・特任助教

研究者番号：３０７０５０５９

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：ストリーミングで与えられた行列を高速で解くことができる非負行列分解のアルゴリズムを導出した．また提案法の性能を評価するため，ソーシャルネットワークサービスの一種であるTwitterに応用を行った．このアルゴリズムを使うことで，日本中の全ツイート（Twitterにおけるメッセージ送信）をリアルタイムで処理することが可能であり，そのため今Twitterでどのような話題が盛り上がっているかを瞬時に知ることが可能であることを確認した．

研究成果の概要（英文）：We proposed a streaming algorithm of non-negative matrix factorization for a sequence of matrices such as time series. We evaluated its performance by applying to the Twitter stream, which is one of the popular social network services. Our algorithm allows to handle all Japanese tweets in real-time, which tells us what topics are "hot" in that moment for every second.

研究分野：　機械学習

キーワード：　非負行列分解　Twitter　トピック抽出　ストリーミング処理

１．研究開始当初の背景

多次元配列あるいはテンソルの構造を持つデータの重要性が高まっている．例えば遺伝子解析，自然言語処理，脳科学など実社会における様々な情報の高次な関係性はテンソルとして表現することができる．また隠れマルコフモデルに代表される隠れ変数モデルの学習にもテンソルが深くかかわっていることが近年知られるようになってきた．従来の学習アルゴリズムの代わりにテンソル分解を使うスペクトラル学習によって高速かつ一意な解を得ることができる．テンソル分解はこれらに共通して必要な基礎技術だが，近年のデータ規模の増加に対応しきれていない．

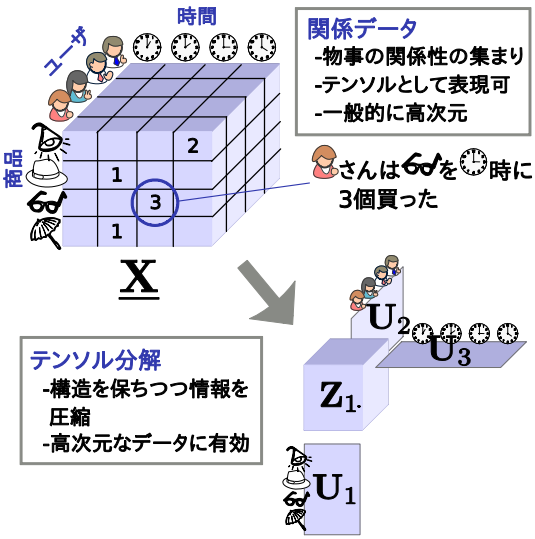

図１．テンソル分解によるデータ表現

２．研究の目的

高速かつ省メモリなテンソル分解のアルゴリズムの開発を行う．というのも，スケーラブルなテンソル分解は現在緊急に求められているからである．

その背景の 1 つに，交通制御，電力網の最適化，スマートシティ構想など実世界に配備されたセンサ群とコンピュータ上の仮想空間を結びつける，いわゆるサイバーフィジカルシステムの存在が挙げられる．これらの事例では実在のセンサ，車，家，人などを互いに関連づけた，従来では考えられない超大規模なテンソルが出現する．その他にも遺伝子解析，自然言語処理，脳科学など多種多様な分野において大規模テンソルが存在する．例えば WEB 上で検索された単語の頻度を保存した Google N グラムは総計 1000 万単語を収録しており，特に連続する 5 単語の出現頻度を表現する 5 グラムデータは 1000 万の 5 という，1 京の 2 乗よりも大きい要素を持つ 5 階のテンソルとして表現できる．この規模の問題が解けるテンソル分解アルゴリズムは現時点で存在しない．

３．研究の方法

本研究ではテンソルデータは本質的には疎性であることに着目し，高速かつ省メモリなテンソル分解のアルゴリズムを開発した．すなわち，テンソルは通常超高次元となるが，意味のある数値（すなわち非ゼロ）が入っている要素数は通常次元に比べると著しく小さいことが多い．この性質を活かし，全要素数ではなく非ゼロ要素数のみに依存するモデルおよびアルゴリズムを導出することで，従来のアルゴリズムよりも大幅に高速化することができる．加えて，非ゼロ要素数に合わせてパラメータの構造を設計することで必要となるメモリ量も大幅に節約することができる．

４．研究成果

（１）期待テンソル分解．具体的にはデータテンソルが何らかの確率変数の期待値として表現できるとき，それをオンラインで解くための枠組みである期待テンソル分解を提案した．この問題では二乗誤差を目的関数とする CP 分解（テンソル分解の一種）を二乗ノルムの正則化付で扱うことができる．また最適化手法として擬似二次情報を使った確率勾配法を導出した．二次の項（ヘッセ行列）をフルに導出するのではなく対角項のみで近似することにより，収束スピードと計算時間の良いトレードオフを達成できた．また収束に関しても理論解析を行い，十分緩い仮定のもと正しい解に行くことが保障される．Amazon レビューデータを用いて性能を評価し，既存のアルゴリズムに比べ，高速かつ省メモリであることを確認した．

（２）Twitter におけるリアルタイムトピック抽出．Twitter に代表される Microblog メディアは，その膨大な使用ユーザ数やリアルタイム性から世の中のトレンドを抽出するための重要な情報源となっている．これらのデータはテキスト文として与えられるため扱いが難しいが，自然言語処理の概念であるトピック抽出によって話題となっているトピックを単語のクラスタとして抽出することができる．

しかしながら，実際に Microblog 上でトピック抽出を行う上で２つの課題が存在する．

① ストリーム性．情報はテキストのストリームとして与えられるため，バッチ処理ではなくオンライン処理を行う必要がある．また実用上トピックの速報性も重要なため，できるだけデータと同じ速度，すなわちリアルタイムで処理を行いたい．しかしながら，そのデータの膨大さのため既存の方法ではこれらを達成できない．

② スパム除去．スパムは特に Twitter において顕著な問題となっている．自動プログラムによるメッセージ送信や広告文

によって同一の文章が多数複製され，これによって無意味なトピックが発生してしまうトピックハイジャックは抽出したトピックの質を著しく損なう．

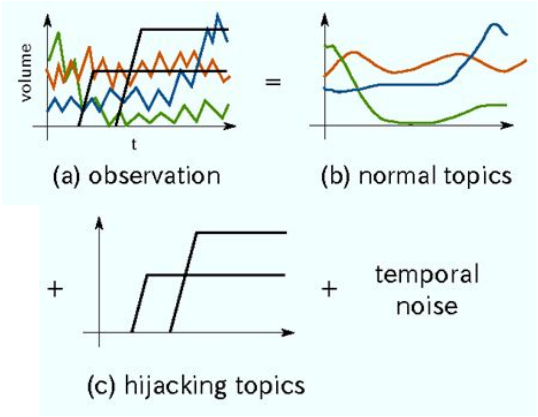

図２．Twitter 上のトピック抽出

我々はこれらの課題を解決するため，期待テンソル分解の枠組みを発展させ，非負行列分解のストリーム学習アルゴリズムおよび因子行列上での統計的検定によるスパムフィルタを提案した．まずデータとして，各時刻毎にテキストから単語の頻度行列（”bag of words”）を観測するとする．この行列を非負行列分解することでトピックが抽出できるが，直接の分解だと時間依存ノイズの影響を受けるため，過去と現在の観測の加重平均の分解を考える．これにより，ノイズに影響されない真のトピックの変動をなめらかに抽出することができ（図 2.(b)），かつ計算量が行列の非ゼロ要素数に対して線形なアルゴリズムが導出できる．

加えて，抽出されたトピックがスパムでないもの（すなわち通常のユーザによって形作られたもの）の場合，そのトピックの単語分布はロングテールになるという，いわゆる Zipf 則が成り立つことを発見した．この性質はスパムからなるトピックには成り立たないものであり，この性質を利用することで自然なトピック（図 2.(b)）とスパムトピック（図 2.(c)）を見分けることができる．具体的には両者を尤度比検定によって比較し，オンラインで除去可能なフィルタシステムを構築した．

計算機実験では約１５００万サンプルからなる実 Twitter データを用い，理論的には日本ユーザの全 Twitter データをリアルタイム処理できることを確認した．さらに実際に抽出したトピックの単語頻度分布と，データと同じ期間に取得した Yahoo! ニュースのヘッドラインに現れた単語を比較し，実世界のニュースイベントと関連性があるトピックが実際に取れたことを確認した．またスパムフィルタによって人工的に挿入したスパム文をほぼ１００％排除できることを確認した．

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　０件）

〔学会発表〕（計　２件）

1. Hayashi, K.; Maehara, T.; Toyoda, M. & Kawarabayashi, K., “Real-time Top-R Topic Detection on Twitter with Topic Hijack Filtering”, ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2015.8.10～2015.8.13（発表確定), Sydney (Australia)

2. 林浩平,「行列・テンソル分解による関係データ解析」, FIT2013 第 12 回情報科学技術フォーラム，2013.9.4～2013.9.6，鳥取大学（鳥取県鳥取市）

〔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：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
林　浩平　　（Kohei Hayashi）
国立情報学研究所　特任助教
研究者番号：30705059

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者

(　　)

研究者番号：