6N-7

テンソル分解を用いたレビューデータの分析

熊本和正† 天笠俊之‡ 丸橋弘治‡ 北川博之‡

†筑波大学情報学群情報科学類 ‡筑波大学システム情報系情報工学域

1 はじめに

近年，大量のデータを分析することによって，新たな知識を獲得する手法が注目され，特にeコマースなどの分野で活発に活用されている．とりわけ，レビューデータは，様々な属性を含むデータであり，これを解析することで，推薦商品の提示や，どのような消費者に人気があるかの調査などが可能になる．一方，3項目以上の関係データを自然に表現し，その上で分析を行うことができるテンソルが注目されている [1]．テンソル上の分析手法の一つとして，テンソル分解がある．これは行列分解をテンソルに拡張したものと理解することができる．テンソルでは，行列では直接扱えない3項以上の関係を自然に扱うことができ，このようなデータにテンソル分解を適用することで，行列分解などでは得られないような構造が解析できることが期待される．本研究では，テンソルデータ化した公開レビューデータを対象にテンソル分解を適用する方法について議論する．

2 前提知識

2.1 テンソル

テンソルは多次元配列である．N次元配列を，本論文ではN次のテンソルと呼ぶ．テンソルは行列を一般化した概念であり，1次のテンソルはベクトル，2次のテンソルは行列となる．また，テンソルの各軸のことを，モードと呼ぶ．また，x_{ijk} は3次のテンソル χ の各モードの添字がそれぞれ (i,j,k) 番目である要素を指す．

2.2 テンソル分解

特異値分解や主成分分析などの行列分解法は，行列形式で表現されるデータの特徴抽出や次元縮約を行う手法として知られている．テンソル分解法は行列分解法を一般化した概念であり，テンソルを低次元のパラメータで近似することができる．テンソル分解法には古典的なPARAFAC分解やTucker分解の他にも，様々な分解法がある．詳しくは [2] を参照されたい．

2.3 PARAFAC分解

PARAFAC分解は，テンソル分解の一種である．PARAFAC分解は，テンソルをランク1テンソルの和に分解する．ランク1テンソルとは，$\chi = a^{(1)} \circ a^{(2)} \circ \cdots \circ a^{(N)}$ のようにN個のベクトルの積で表現可能なN次のテンソル $\chi \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ のことを指す．任意のテンソル $\chi \in \mathbb{R}^{I \times J \times K}$ にPARAFAC分解を適用すると，各要素は

$$x_{ijk} \approx \sum_{r=1}^{R} a_r b_r c_r (\text{ただし}, i = 1, ...I, j = 1, ...J, k = 1, ...K) \quad (1)$$

で近似される．ここで，行列 $A \in \mathbb{R}^{I \times Q}$,$B \in \mathbb{R}^{J \times Q}$,$C \in \mathbb{R}^{K \times Q}$ は，それぞれ分解されるテンソル χ の各モードに対する低次元表現と解釈することができる．

3 テンソルを用いたレビューデータ分析法

3.1 概要

レビューデータには様々な情報が含まれるが，本研究では,「どの時期に」,「どのユーザが」,「どの施設に」レビューを書いたかという情報を用いて，テンソルデータを作成する．このテンソルデータにPARAFAC分解を適用した結果から，施設のクラスタリングを行う．クラスタリングを行うことで，施設間の隠れた関係性を明らかにすることができると期待される．

3.2 手順

3.2.1 データセットからテンソルデータへの変換

レビューデータから「投稿日時」,「ユーザ名」,「施設名」を抽出し，各々1からはじまる「週ID」,「ユーザID」,「施設ID」を割り当てる.「ユーザID」には (1〜ユーザ数),「施設ID」には (1〜施設数) のユニークなIDを割り当てる.「週ID」には，その年の何週目か (1〜53) を割り当てる．たとえば，1月第2週にユーザID=25のユーザが施設ID=6の施設に投稿したレ

Analyzing User Reviews by Tensor Decomposition

Kazumasa KUMAMOTO†(kumamoto@kde.cs.tsukuba.ac.jp), Toshiyuki AMAGASA‡(amagasa@cs.tsukuba.ac.jp) and Koji MARUHASHI‡(kojimaru@kde.cs.tsukuba.ac.jp) and Hiroyuki KITAGAWA‡(kitagawa@cs.tsukuba.ac.jp)

†College of Information Sciences, University of Tsukuba

‡Faculty of Engineering, Information and Systems, University of Tsukuba

ビュー」は $x_{2,25,6} = 1$ に変換される．投稿がない要素については 0 とする．

3.2.2 テンソル分解

前項で作成したテンソルデータについて，PARAFAC 分解を適用する．ここで，レビューデータにはスパムユーザや誤送信による大量の同内容の投稿が含まれている場合がある．このようなノイズデータを含んだデータにテンソル分解を適用すると，分解の結果がノイズに影響され，意図した結果が得られない．このため，ノイズデータはデータからあらかじめ除去する．具体的には，PARAFAC 分解の結果，極端に高いスコアを示すユーザの投稿を，テンソルデータから除去し，除去されたテンソルについて再度テンソル分解を行う．

3.2.3 クラスタリング

PARAFAC 分解の結果，各施設について指定した次元数のスコアベクトルを各々抽出できる．これらについて，ユークリッド距離に基づいて群平均法 (UPGMA) でクラスタリングする．

3.2.4 ジオコーディング

クラスタリングされた施設がどのような位置に分布するのかを調査するため，レビューデータに位置情報が含まれない場合，Yahoo!ジオコーダ API[1] を使用して位置情報を取得する．ジオコーダ API は住所あるいは施設名を指定すると，緯度，経度などの情報を返す．

4 実験

本研究では，楽天株式会社が公開する，楽天トラベルの公開レビューデータを用いる [2]．レビューデータにはレビューの「投稿時間」，マスクされた「ユーザ名」，レビュー対象となる「施設名」などが含まれる．位置情報の取得に施設名を使うと，一意に定まらない可能性があるため，楽天トラベル [3] の詳細情報ページから住所を取得して，施設名の代わりに住所を指定した．PARAFAC 分解の適用には，数値解析ソフト MATLAB R2013a および MATLAB Toolbox[4] を使用した．

4.1 実験手順

まずデータセットをテンソルデータに変換し，PARAFAC 分解を適用する．この際使用するレビューデータは，投稿回数は 8〜20 回のユーザの投稿のみ，2011 年の投稿のみに制限した．この結果極端に高いスコアを示すユーザの投稿を除去し，再度 PARAFAC 分解を適用する．分解で得られた施設ごとのベクトルから，施設について UPGMA でクラスタリングを行う．

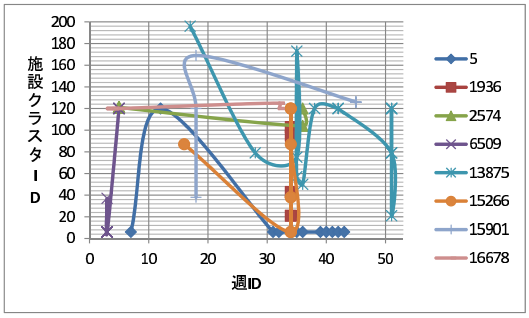

図 1: クラスタごとのユーザレビューの傾向

4.2 実験結果

図 1 はユーザを 200 のクラスタに分類した結果のうち，クラスタ番号 31 の結果を示している．横軸は時間（週 ID，1〜53），縦軸は施設クラスタ ID（1〜200）であり，各ユーザ投稿したレビューの時系列（週 ID）の昇順に，どの施設にレビューを書いたかによって軌跡をプロットしている．これから，複数のユーザが似たような時期（夏季）に集中してレビューを投稿していることがわかる．またこれらのユーザが評価している施設は，地理的にはばらついていることが，位置情報から判明した．このように，ユーザ，施設および時間を考慮したレビューデータの分析が可能になる．今後は，さらに詳細な分析を行う予定である．

5 まとめ

本論文では，投稿時間，ユーザ名，施設名から成るテンソルに対して PARAFAC 分解を適用し，施設のクラスタリングを行った．今後の課題としてはこのクラスタリング結果についての詳細な考察と，レビューデータの評価値など，他の項目をを加えた分析などが挙げられる．

謝辞

貴重なデータを提供していただいた楽天技術研究所に深い感謝の意を表する．

参考文献

[1] Panagiotis Symeonidis. User Recommendations based on Tensor Dimensionality Reduction Artificial Intelligence Applications and Innovations III, 2009.

[2] T. G. Kolda and B. W. Bader. Tensor decompositions and applications Technical report, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, December 2007.

[1] Yahoo!ジオコーダ API http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/geocoder.html

[2] 楽天データ公開 http://rit.rakuten.co.jp/rdr/

[3] 楽天トラベル http://travel.rakuten.co.jp/

[4] Brett W. Bader, Tamara G. Kolda and others. MATLAB Tensor Toolbox Version 2.5, January 2012. http://www.sandia.gov/~tgkolda/TensorToolbox/