F-001

グラフマイニングに基づくアイテム推薦における負の評価値の利用法
Using Negative Ratings in a Graphmining-based Recommender System

松井 淳　山田 一郎　藤井 真人　苗村 昌秀
Atsushi Matsui　Ichiro Yamada　Mahito Fujii　Masahide Naemura

1. まえがき

日々大量のデータが生成され発信される情報化社会では，膨大なデータの中から有益な情報を自動的に抽出する技術が求められている．その一方で，現実に観測可能なデータセットには相当数の欠損値が含まれている場合が多く，ビッグデータ解析の一分野であるアイテム推薦における研究課題となっている．

このような大規模データ解析における疎性の影響を緩和する一つの方法としてグラフマイニング（Random Walk with Restart, RWR[1]）が提案されており，近年，注目を集めている．しかしグラフマイニングでは，解析対象のグラフのエッジに付与される重みは，エッジが結ぶノード間の遷移確率にあたる非負のパラメータであるため，エッジ重みの根拠となるアイテム評価値が正負の二方向の尺度で観測された場合であっても，両者の値を同一のグラフに反映させることは原理的に難しい．従って，従来の報告例の多くは，負の評価値を無視した条件でグラフを作成し解析してきた．しかし，負の評価値は推薦対象として望ましくないアイテムに関する情報（負の嗜好）を保持していると考えることができるため，正・負の評価値の全てを解析対象とすることが望ましい．

本報告では，番組評価値のデータセットを用いた予測実験を通じて，グラフマイニングにおける負の評価値の利用法について考察する．

2. グラフマイニングに基づくアイテム推薦

2.1 基本アルゴリズム

グラフマイニングの基本更新式を(1)式に示す．ベクトル $\boldsymbol{p}(t)$は，個々のユーザ，アイテム，および，アイテムの属性を記したメタデータ（単語）をあらわす個々のノードの第 t ステップ目における滞留確率である．ベクトル $\boldsymbol{q}$ は，探索の起点となるノード（被推薦ユーザ）の滞留確率のみが 1 の初期ベクトルである．隣接行列 A は，(2)式に示すように，9 つの部分行列で構成される．部分行列 IU の各要素は，観測されたアイテム評価値によって定まるユーザ・アイテム間の遷移確率である．グラフマイニングは，ベクトル $\boldsymbol{p}(t)$を一定の確率αで初期ベクトル q に再初期化しながら定常状態に到達するまで(1)式の更新を繰り返し実行することによって，部分行列 IU に記載されていない欠損部分の値（すなわち，未知の評価値の予測値）を算出する．ここで，MI はアイテム毎のメタデータの出現関係を記した二値の行列であり，MM はメタデータ間の意味的関係を記した類似度行列である．また UI, IM は，それぞれ IU, MI の転置行列である． UU, II, UM, MU は全て零行列とする．なお，隣接行列 A は，各列の和が 1 となるように正規化されているとする．

$$\boldsymbol{p}^{(t+1)} = (1-\alpha)A\boldsymbol{p}^{(t)} + \alpha\boldsymbol{q} \tag{1}$$

$$A = \begin{pmatrix} UU & UI & UM \\ IU & II & IM \\ MU & MI & MM \end{pmatrix} \tag{2}$$

2.2 二極性グラフマイニング

隣接行列 A の部分行列 IU が定めるユーザ・アイテム間の遷移確率は非負値である必要がある．従来のグラフマイニングでは，IU の根拠となるアイテム評価値が正負二極の枠組みで与えられる場合であっても，一方の値しか利用することができない．しかし，負の評価値は推薦対象として好ましくないアイテムに関する情報（負の嗜好）を反映した観測データであるため，それらを全て無視することは嗜好に関する情報の損失をもたらす．我々は，アイテム評価値のデータをその符号によって 2 つの集合に分割し，それぞれを独立に解析する二極性グラフマイニング[2]を考案した．この手法は，正の評価値から得られる positive score: S_{POS} と，負の評価値から得られる negative score: S_{NEG} の両者を用いてアイテム毎のスコア（item score）を計算する．

二極性グラフマイニングによる解析結果の具体例として，放送済番組に対する評価値（詳細は次節を参照のこと）についての，正のサンプル（評価値: +2）と負のサンプル（評価値: -2）の S_{POS}・S_{NEG} 平面上での散布図を**図 1** に示す．図 1 より，推薦リストに入れたい正のサンプルの分布と推薦リストから除外したい負のサンプルの分布は重なりを持つが，両者の比 R が大きい領域（図 1 左上）では，R が小さい領域（図 1 右下）よりも正のサンプルが相対的に多く出現している様子がわかる．

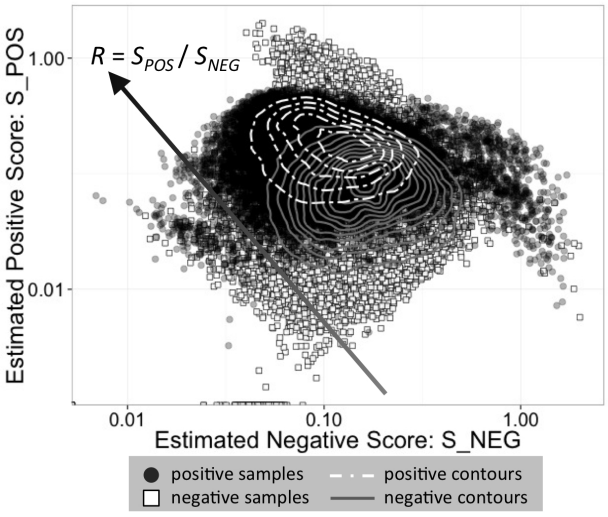

図 1　二極性グラフマイニングの結果の散布図

NHK 放送技術研究所 (ハイブリッド放送システム研究部)

3. 評価値予測タスクにおける負の評価値の利用法

文献[2]に記した二極性グラフマイニングの実装形態では，アイテムの推薦順位を決定する *item score* を，個々のアイテムごとに算出した二種類のスコア S_{POS} と S_{NEG} の差分で定義しているが，この算出方法が最適である理論的な保証はない．また，従来のグラフマイニングや，二極性グラフマイニングの個々の計算（正・負の評価値別の解析）では，評価値のデータセットを符号によって２つに分割し，一方のデータセット（負の評価値については，その絶対値）のみを用いて行列 *UI* を作成するが，*UI* の各要素が対応する評価値の間の大小関係が保持されていれば，原理上は単一のグラフマイニングによる解析が可能である．従って，参照する評価値の尺度が非負となるような変換をあらかじめ施したならば，負の評価値を利用するためにデータセットを分割する必要はない．本報告では，*item score* の具体的な定義や，行列 *UI* の作成方法に関する４つの異なる手法を示し，それぞれの手法による推薦結果の精度を，番組評価に関するデータセットを用いた交差検定法によって比較・検証する．

本報告で取り上げる４つの解析方法の名称と概要を**表１**に示す．RWR_pos は，与えられたデータセットの正のデータのみを用いて解析した結果（S_{POS}）をそのまま *item score* として採用する． RWR_shift は，与えられたデータセットの最小値が 0 となるようなバイアス項を全てのデータに加え，評価値全体が非負となるような変化を施した上で通常のグラフマイニングによる解析を行う．RWR_diff は，二極性グラフマイニングによって得られる二種類のスコア S_{POS} と S_{NEG} の差分を *item score* とする．RWR_ratio は，S_{POS} と S_{NEG} の比率で *item score* を定義する

実験で用いるデータセットは，NHK 放送済番組 300 本に対する 1228 人の被験者の評価値（5 段階評価，+2：見たい ～ -2：見たくない）を用いた．評価値の予測結果の評価基準は，*item score* の上位 *k* 番組からなる推薦リストの正解率 *Precision@k* を用いた．ただし，ここでの「正解」は，真の評価値が+2 または+1 の番組とした．

評価値の一部をランダムに欠損させた場合の，データ欠損率に対する予測精度の変化の様子を**図 2** に示す．図２に示した結果より，データ欠損率によらず，負の評価値を独立に解析し，正の評価値の解析結果と統合する手法（RWR_diff, RWR_ratio）の方が，正の評価値のみを解析する手法（RWR_pos）よりも高い予測精度を示した．一方，全ての評価値が正の値となるようにシフトした上で RWR によって解析した場合の予測精度（RWR_shift）は，今回実施した４通りの解析方法の中で最も低かった．これは，評価値をシフトすることによって，「見たい」番組と「どちらでもない」番組に対する各々の遷移確率のダイナミックレンジが狭まった結果，「見たい」番組についての *item score* と，その他の番組に対する *item score* との間に差が生じにくくなったためであると考えられる．

次に，欠損率 0%の設定における推薦リスト長 *k* と予測精度 *Precision@k* との関係を**図 3** に示す．図３に示した結果より，推薦対象の番組を予測評価値の下位にまで拡大した場合であっても，S_{POS} と S_{NEG} の二種類のスコアを独立に解析し統合する手法の方が，単一のグラフマイニングで解析する方法よりも高い予測精度を示した．

S_{POS} と S_{NEG} の統合方法については，図２，図３のグラフが示すように，両者の差を評価する手法（RWR_diff）よりも，両者の比を評価する手法（RWR_ratio）の方が総じて高い予測精度を示した．

表１．負の評価値を含むデータセットの解析方法

RWR_pos	S_{POS} のみを解析し評価（S_{NEG} は無視）
RWR_shift	評価値全体を正にシフトして S_{POS} を解析
RWR_diff	S_{POS} と S_{NEG} の差を評価
RWR_ratio	S_{POS} と S_{NEG} の比を評価

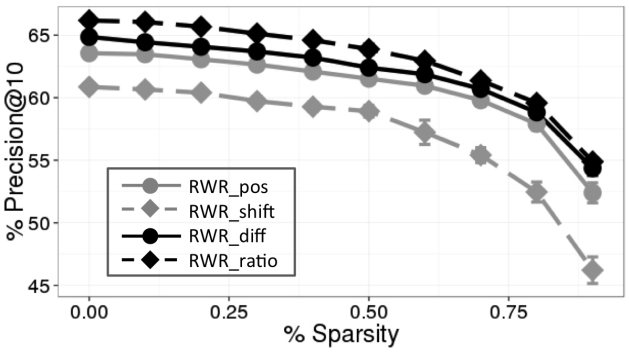

図２　データの欠損率と予測精度（Precision@10）

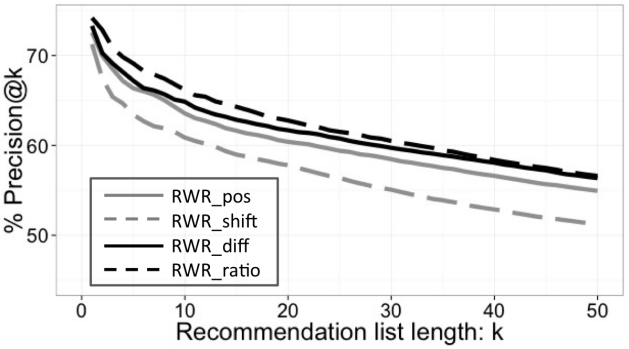

図３　推薦リストの長さと予測精度（Precision@k）

4. まとめ

データの希薄問題を回避可能なグラフマイニングにおいて，正負二極の評価値が観測可能である場合の，負の評価値の利用方法について検討した．NHK 放送済番組 300 本に対する評価値のデータセットを用いた比較実験を行った結果，今回検討した４つの解析方法の中では，正・負の評価値それぞれから得られる二種類のスコアの比を評価する方法が，データの欠損率や推薦リストの長さの条件によらず，最も高い予測精度を示した．

参考文献

[1] H. Yildirim, and M. S. Krishnamoorthy, “A random walk method for alleviating the sparsity problem”

[2]松井淳, 藤井真人, 苗村昌秀, “二極性グラフマイニングによるテレビ番組推薦”, 通信総大講演論文集, D-8-9 (2012).

[3] 松井淳, 宮崎太郎, 山田一郎, 藤井真人, 苗村昌秀, “多段グラフマイニングによる新規アイテム推薦”, 映情技報, Vol.37, No.20, ME2013-61 (2013).