

近傍エッジとの関係に着目したグラフマイニング手法の提案と評価

Proposal and Evaluation of Graph Generation Method

Based on Relative Value of Neighbor Edges

稲福 和史[†] 伏見 卓恭^{††} 佐藤 哲司^{†††}

[†] 筑波大学情報学群知識情報・図書館学類 〒305-8550 茨城県つくば市春日 1-2

^{††} 東京工科大学コンピュータサイエンス学部 〒305-0982 東京都八王子市片倉町 1404-1

^{†††} 筑波大学図書館情報メディア研究科 〒305-8550 茨城県つくば市春日 1-2

E-mail: [†]{inafuku,satoh}@ce.slis.tsukuba.ac.jp, ^{††}takayasu.fushimi@gmail.com

あらまし 近年、EC サイトにおける購買行動履歴など、様々な行動履歴データをグラフとして表現・分析し、商品等の推薦に利用するグラフマイニングでの応用が広がっている。一般に、グラフマイニングでは高頻度に出現する構造を対象に分析を行う。しかし、共購買ネットワーク等のグラフでは、高頻度でなくても局所的な構造が重要となる場合がある。本研究では、グラフの局所的な構造に着目したグラフマイニング手法を提案する。提案法では、局所的な構造の抽出を目的に、隣接エッジ集合の重みから Z スコアを算出、Z スコアが閾値以上のエッジを分析対象とする。実運用されている EC サイトのレビューデータを購買データとみなし、ユーザの購買順序から有向グラフを構築、提案法の有効性を検証したところ、高い Z スコアを示すエッジから重要な構造が抽出できることが明らかになった。

キーワード データマイニング、グラフマイニング、オンラインショッピング、購買行動

1. はじめに

近年、携帯型インターネット接続端末の普及や計算機の性能向上、ストレージの低廉化などを背景として、様々なサービスにおいて、大量のユーザ行動履歴データの取得・蓄積が可能となった。例えば、EC サイトにおける購買履歴、インターネットテレビにおける視聴履歴、観光地を巡る観光者の移動履歴などが挙げられる。

これらのデータは、商品や番組、観光地をノード、ノード間の遷移関係をエッジ、遷移回数を重みとして扱うことで、重み付き有向グラフで表せる。本研究では、購買履歴グラフ (Purchase History Graph) (以下、PHG) を対象に議論を展開するが、視聴履歴や移動履歴など、順序性を備えたグラフであれば適用可能なグラフマイニング手法の提案を目指す。

大量の時系列データからグラフ構造を構築し、有用な知見となりうる特徴的な構造を抽出するグラフマイニング研究が盛んである [3], [5], [9], [11] [4]. 「有用」の捉え方には様々なアプローチがあるが、最も主流な手法として、頻出する構造は有用な知見となりうる、という考え方に基づく頻出パターンマイニングがある。グラフ全体から、頻出する一部の構造やノード、エッジを抽出すること、例えば、PHG からは、典型的な購買パターンや連続して購入されやすい商品の組み合わせを抽出することで、売上向上や顧客満足度向上に有効な知見を得られるとする考え方である。

グラフを大域的に捉え抽出対象を決定する従来手法とは異なり、本研究では、グラフの局所性に着目したグラフマイニング手法を提案する。グラフの局所性について説明する。スマー

トフォンと関連アクセサリーの共購買ネットワークの模式図を図 1 に示す。エッジの太さはエッジの出現回数 (商品の共起回数) を表す。このグラフから、特に太いエッジを抽出することで特定のスマートフォンに人気のあるアクセサリーを知ることができる。図 1 について、出現回数の多いエッジに着目すると、左側のスマートフォンが青いケースと緑のストラップと同時に購入される傾向にあることが分かる。しかし、右側のスマートフォンの関係に着目すると、こちらからも緑のストラップと同時に購入される傾向を観察できる。この時、出現回数を基準に分析対象を選別、あるいは絞ると「右側のスマートフォンと緑のストラップ」の関係が失われてしまう。

このように局所的な抽出を行わなければ、得られない知見が存在する。同様に、現実世界を対象として生成されるグラフは、極めて局所性の高いグラフとなることが多いと考えられる。これに対し、本研究では、近傍エッジとの関係に着目しエッジの抽出を行うグラフ生成手法を提案する。提案手法により、共購買ネットワークのような局所的な構造を持つグラフの高頻度な部分グラフや低頻度な部分グラフにおいて、それぞれの部分グラフの特徴を考慮したエッジの抽出が可能になる。

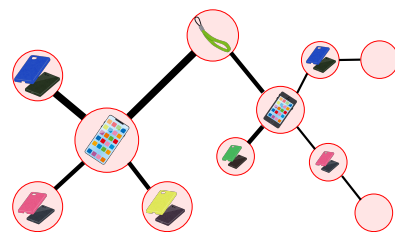


図 1 共購買ネットワークの例

本論文の構成は次のとおりである。まず、関連研究について 2. 章で説明し、次に近傍エッジとの関係に着目したグラフ生成手法について 3. 章で説明する。続いて、4. 章で実運用されている EC サイトのデータセットに対し、提案手法を適用し、グラフマイニングを行った。5. 章で実験結果についての考察を行い、最後に 6. 章で本研究についてまとめている。

2. 関連研究

本研究で提案する手法は、グラフの局所構造に着目した分析手法といえる。局所構造の抽出は、コミュニティの抽出と類似性が高い。小さな幾つかの集まりが連結してできたグラフでは、それらをコミュニティ毎に分割して分析を行うなどの手法が一般的である。

コミュニティ分割の指標としてモジュラリティが広く使われているが、Fortunato らは小規模なクラスタが多く含まれるようなデータにおいては、モジュラリティが有効に機能しないことを示している [2]。また、chen らは、大規模なコミュニティの中には、多くのノイズが混じっており、このノイズを取り除くことで、より実態に沿ったコミュニティを抽出できるとして、新たな中心性指標及びコミュニティ検出手法を提案している [1]。

このように、小規模なクラスタや局所構造の抽出を試みた研究はいくつかあるが、これらは無向・重みなしグラフを対象としている。

本研究では、提案手法によって生成したグラフの評価には、購買行動パターンの抽出を用いる [16]。この手法の核となるのは、アイテムからなるネットワークのモチーフ分析である。モチーフ分析、アイテムベースのネットワーク分析及びネットワークベースのアイテム推薦に関する関連研究について説明する。

Salihoglu は無向ネットワークを 3 ノード及び 4 ノードのモチーフパターン出現頻度ベクトルで表現し、道路ネットワーク、AS ルーティングネットワーク、共購買ネットワークの構造はモチーフパターン出現頻度ベクトルに反映されることを明らかにしている [8]。Srivastava はモチーフパターンの出現頻度とコミュニティ検出によって Amazon の共購買ネットワークを分析している [10]。著者はネットワーク全体に対するモチーフ分析を行ったのち、ネットワーク全体をコミュニティに分割、ノードとして扱い新たなネットワークを構築、モチーフ分析を行った。これにより、さらに俯瞰的な視点での分析が可能になることを示した。また、ネットワーク全体及びコミュニティをノードとしたネットワークにおいて、同様のモチーフパターン出現頻度が 3 ノード及び 4 ノードの両方から得られたことを明らかにしている。Zhou はアイテムとユーザから構築される二部グラフを用いたネットワークベースのアイテム推薦を提案している [15]。また、人気の高いアイテムが極端に高いスコアを獲得することを防ぐため、スコアの平均化プロセスを追加した [14]。しかし、この方法は人気の低いアイテムに過度のスコア補正を与えてしまうため、改善手法が提案されている [6], [7]。また、Yu によって多数のネットワークベースのアイテム推薦手法が収集・比較されている [13] が、上述の手法を含め、これらも全

て無向グラフに基づくものである。

3. 近傍エッジとの関係に着目したグラフ生成手法

本研究では、EC サイトにおける個々のユーザの購買順序から購買順序グラフ (以下、PHG) を構築し、隣接エッジ集合内のエッジ重みの相対値を算出、閾値以上の相対値を示すエッジを分析対象としてサブグラフを生成する手法を提案する。

ユーザ集合及びアイテム集合をそれぞれ U 及び I と定義する。ユーザ u がアイテム i のレビューを時刻 t に投稿したとき、レビュー r を $r = (u, i, t)$ と表す。ユーザ u が N 件のレビューを投稿した際のレビュー群を $R(u) = [r_1, r_2, \dots, r_{N_u}]$ と表す。 r_n は r_{n+1} より 1 つ前に投稿されたレビューであることを示す。ここで、ユーザ u にレビューされたアイテム群を $I(u) = [i_1, i_2, \dots, i_{N_u}]$ と定義し、 $I(u)$ において、連続する要素をペアにした集合を構築する。

$$SI(u) = \{(i_1, i_2), (i_2, i_3), \dots, (i_{N_u-1}, i_{N_u})\} \subset I(u) \times I(u).$$

得られた SI を基に PHG を構築し分析する手順を以下に示す。

3.1 購買順序グラフの構築

各ユーザのレビューデータから有向グラフを構築する。3 名のユーザが合計で 5 種類、6 個のアイテムを購入した際に構築される PHG の模式図を図 2 に示す。アイテムをノードと定義すると、ノード集合 V はアイテム集合 I と等しいため、 $V = I$ である。ここで、 $SI(u)$ のノードペア (i, j) 間に i から j への有向エッジを追加すると、エッジ集合 E は次のように表せる。

$$E = \left\{ (i, j) \in \bigcup_{u \in U} SI(u) \right\} \quad (1)$$

この $G = (V, E)$ で表される有向グラフを PHG とする。また、ノードペア (i, j) の出現回数をエッジの重み $w(e(i, j))$ とする。

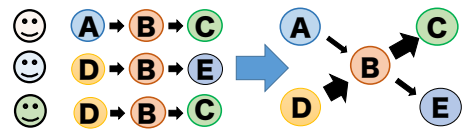


図 2 購買順序グラフ構築の模式図

3.2 隣接エッジ集合内の重みに基づく相対値の算出

隣接エッジ集合内の重みから相対値を算出する。相対値の算出には、偏差値の計算などに用いられる Z スコアを採用する。エッジ $e(i, j)$ の隣接エッジ集合を $\Gamma(e(i, j)) = \{(i, k) \in E\} \cup \{(k, i) \in E\} \cup \{(j, k) \in E\} \cup \{(k, j) \in E\}$ と定義する。この時、エッジ $e(i, j)$ の Z スコア $Zscore(e(i, j))$ は次のように計算される。

$$Zscore(e(i, j)) = \frac{w(e(i, j)) - \mu(e(i, j))}{\sigma(e(i, j))},$$

ここで、 $\mu(e(i, j))$ 及び $\sigma(e(i, j))$ は、それぞれ、エッジ $e(i, j)$ の隣接エッジ集合 $\Gamma(e(i, j))$ の重みの平均と標準偏差を表して

おり、 $\mu(e(i, j)) = \frac{1}{|\Gamma(e(i, j))|} \sum_{e \in \Gamma(e(i, j))} w(e)$ 及び $\sigma(e(i, j)) = \frac{1}{|\Gamma(e(i, j))|} \sum_{e \in \Gamma(e(i, j))} w(e)^2 - \mu(e(i, j))^2$ のように計算される。

3.3 サブグラフの抽出

3.1 及び 3.2 で説明したように、PHG のエッジには出現頻度と Z スコアの 2 つの重みが付与されている。これらを用いて、2 種類のサブグラフを生成する。1 つ目は、出現頻度 $w(e(i, j))$ が閾値 θ_w を超えたエッジ集合から構成されるグラフである。本研究ではこれを Conventional-PHG と呼ぶ。2 つ目は、Z スコア $Zscore(e(i, j))$ が閾値 θ_z を超えたエッジ集合から構成されるグラフである。本研究ではこれを Proposed-PHG と呼ぶ。

4. 実験と評価

4.1 データセット

本研究では、Nii が研究用に公開している EC サイト「楽天市場」のレビューデータ^(注1)を使用した。これは、レビューの投稿順序は購買順序を表すとみなせるためである。6,500 万レビューからなるデータセットから、投稿者が一意に判別できるレビューを抽出した。その際、購入した事が確認できない、あるいは投稿日時が欠落しているレビューを除外した。以上の処理を行い、2,445,084 ユーザによる 17,794,337 レビューを実験対象のレビュー集合 $R = \bigcup_{u \in U} R(u)$ とした。

4.2 購買順序グラフの構築

4.1 のデータから、6,721,476 ノード、24,325,221 エッジの PHG を構築し、Z スコアを算出した。エッジの出現頻度及び Z スコアの散布図を図 3 及び図 4 に示す。横軸が出現頻度、縦軸が Z スコア、一つの点が一つのエッジに対応している。従来手法では、出現頻度が閾値以上を示すエッジを抽出し分析を行ってきた。一方、提案手法では Z スコアを基準に抽出対象のエッジを決定する。

従来手法により、図 3 に示す緑の枠内の点から構築されるサブグラフを Conventional-PHG、提案手法により図 4 に示す赤い枠内の点から構築されるサブグラフを Proposed-PHG と呼ぶ。本研究においては、出現頻度の閾値 θ_w を 10、Z スコアの閾値 θ_z を 12 とした。Z スコアの閾値は 4. で説明する、購買行動パターンの抽出において、Conventional-PHG と Proposed-PHG の弱連結成分数になるべく近くなるように設定した。

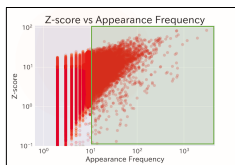


図 3 従来手法の
分析対象エッジ

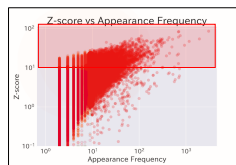


図 4 提案手法の
分析対象エッジ

4.3 購買行動パターン分析

4.2 で構築した 2 グラフから、4.3.1 に示すモチーフ分析を軸とした購買行動パターンの抽出を行い、それぞれのグラフの有する特徴について分析を行う。多くのユーザの購買順序から構築する PHG からは、商品の買われる順番に着目した知見を得ることができる。PHG を弱連結成分 (以下、WCC) に分解し、WCC 毎に特徴量を求め、クラスタリングを行う。ここで、各クラスターのセントロイドを典型的な商品の買われ方、すなわち購買行動パターンとみなす。以下に、各ステップについて詳細に説明する。

4.3.1 購買行動パターンの抽出

a) 弱連結成分分解

4.2 で構築した 2 つのグラフに対し、弱連結成分分解を行う。これは、一つの弱連結成分が一連の購買行動を表すと考えられるためである。図 5 のように、グラフ $G(V, E)$ において、ノード u とノード v が u から v ないし v から u に到達しうるとき、 u と v は同じ弱連結成分 (WCC) に属する。ここで、WCC 数を M とし、第 m 番目の WCC に属するノードを $V^{(m)}$ 及び $E^{(m)}$ と定義すると、各 WCC グラフは $G^{(m)} = (V^{(m)}, E^{(m)})$ で示される。本研究においては 5 ノード以上からなる WCC を抽出対象とした。

b) モチーフベクトルの構築

WCC にモチーフ分析を適用する。3 ノードからなるモチーフパターン (MP) は 13 種類あり (図 6)、WCC $G^{(m)}$ にモチーフ分析を行うことで、モチーフ分布 $\mathbf{x}_m$ が算出できる。 $\mathbf{x}_m$ は各 MP の出現頻度を表しており、13 次元ベクトルとなる。このベクトルを正規化 (L_1 norm) しモチーフの出現確率ベクトルを取得する。これをモチーフベクトル^(注2)と称する。

グラフの構造をモチーフベクトルを用いて表現することで、より直感的に WCC の特徴を捉えることができる。例えば、モチーフパターン 1(MP1) は 1 種のメイン商品から多種のオプション商品への購買行動を示していると考えられる。具体的には「スマートフォンとスマートフォンケース」のような関係である。MP2 は購買される順序が決まっている商品を示しており、「コミックの 1 巻、2 巻、3 巻」というようなケースが考えられる。また、MP13 は購買の順序性がない「服」などの購買行動を示していると考えられる。基本となる購買行動パターンは 1 つの MP で表せることが望ましい。しかし、実際の購買行動パターンは複数の MP の組み合わせ (混合モデル) になると考えられる。そのため、WCC の構造的特徴をモチーフベクトルを用いて表現する。

c) k -means クラスタリング

構築したモチーフベクトル $\mathbf{x}$ に k -means クラスタリングを適用し、WCC を K 個のクラスターに分ける。WCC $G^{(m)}$ と $G^{(n)}$ の距離は次のように求める。

$$d(\mathbf{x}_m, \mathbf{x}_n) = \|\mathbf{x}_m - \mathbf{x}_n\|_{L_2}^2. \quad (2)$$

各クラスターはモチーフベクトルに共通の傾向を持つ WCC に

(注1) : <http://www.nii.ac.jp/dsc/idr/rakuten/rakuten.html>

(注2) : モチーフパターン出現頻度の累積度からより典型的な購買行動パターンを抽出するため、モチーフ分布ベクトルを確率ベクトルへと変換している。

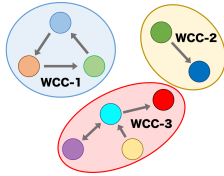


図5 弱連結成分分解

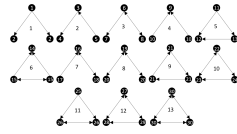


図6 3ノードモチーフ

よって構成される。すなわち、一つのクラスタは一つの典型的な購買行動パターンであると考えられる。ここでは、各クラスタにおけるセントロイドのモチーフベクトルを代表値として抽出、中心となるMPの組み合わせから、購買行動パターンとしてどのような意味を持つのか分析する。クラスタ数の決定はGAP統計量 G_k [12] を用いた。 G_k を最小とする K が求めるクラスタ数となる。

d) 主成分分析

構築したモチーフベクトルに対し、主成分分析を行う。上位の主成分の固有ベクトルを求め、その固有ベクトルについて上位及び下位の要素を観察することで、購買行動パターン間の関係を決定する要因となるMPを抽出・分析できる。

具体的には、各WCC $G^{(m)}$ の13次元モチーフベクトル $\mathbf{x}_m$ を集めた $M \times 13$ の行列 $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_M]^T$ から 13×13 の共分散行列

$$\Sigma = (\mathbf{X} - \mu)^T (\mathbf{X} - \mu)$$

を計算し、 Σ の固有ベクトル群 $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_{13}]^T$ を求める。ここで、 $\mu = [1, \dots, 1]^T [\mu_1, \dots, \mu_{13}]$ は、各モチーフパターンの出現頻度の平均を要素とした $M \times 13$ の行列である。値が大きい H 個の固有値 $\lambda_1, \dots, \lambda_H$ と対応する固有ベクトル $\mathbf{v}_1, \dots, \mathbf{v}_H$ を抽出する。

4.3.2 Conventional-PHGの購買行動パターン分析

a) k -means クラスタリング

Conventional-PHGに対し、弱連結成分分解を行った。その結果、141WCC、2,094ノード及び5,399エッジからなるグラフが構築された。この141WCCからモチーフベクトルを取得し、 k -means クラスタリングを適用した。GAP統計量 G_k により、クラスタ数は8とした。表1は、クラスタ別のWCC数を示す。また、各クラスタのセントロイドのモチーフベクトルを図7に示す。

表1 クラスタ別 WCC 数 (Conventional-PHG)

Cluster-1	Cluster-2	Cluster-3	Cluster-4
17	8	14	38
Cluster-5	Cluster-6	Cluster-7	Cluster-8
28	11	19	6

各クラスタの特徴について、MPの構造を軸に説明する。

• Cluster-1

モチーフベクトルに着目すると、もっとも多く出現するMPがMP2、続いてMP3、MP7である。これらはいずれも2ノード間の関係から構築されるパターンであり、特にMP2はコアノードを経由する一方通行型である。このパターンにおいて

ユーザは、コアノードを経由して購買行動を次々と発展させていると考えられ、Cluster-1は「中心経由発展型」であると考えられる。

• Cluster-2

本クラスタでは、MP7が最も多く観察され、第2位にMP2、第3位にMP4が続く。これらのMPはいずれも1ノードへ集中する傾向を持つ。このことから、Cluster-2はコアノードとその周辺ノードによって主に構成され、コアノードへ集中するパターンであると考えられる。具体的には、多種多様なメイン商品群とそれらに共通して使用可能なオプション商品の組み合わせが想定される。

• Cluster-3

Cluster-3はMP8、MP3が極端に多く出現するクラスタであり、いずれのMPも2ノード間の関係からなる。Cluster-2と同様にコアノードと周辺ノードによって構成されるクラスタであるが、その関係性が双方向エッジであるという点で異なり、コア商品と周辺商品の前後を問わず購買が行われる。またコアノードは始点にも終端にもなることから、コアノードが購買行動を強く促す役割を持つといえる。

• Cluster-4

Cluster4は全クラスタの中で最大のクラスタであり、非常に複雑なモチーフ分布を示した。WCC毎のノード数・エッジ数の標準偏差はそれぞれ116と363と他のクラスタと比較して大きな値を示した。cluster4は最大連結成分(LWCC)を有しており、これが標準偏差に大きな影響をもたらしていると考えられる。そのため、LWCCを除外してノード数・エッジ数の標準偏差を算出したところ、17と50であった。結果として、LWCCの影響を差し引いてもほかのクラスタに比べ大きな標準偏差を示した。以上から、Cluster4は様々なWCCを含む複雑型であると考えられる。

• Cluster-5

Cluster-5は主にMP3とMP1から構成されるクラスタであり、いずれのMPもコアノードから周辺ノードへの有向エッジが多い、放出型MPである。そのため、Cluster-5では、コアノードが購買行動の起点となる。また、さらなる連続購買へと発展せず、短いシーケンスで完結する傾向にある。Cluster-5は1種類のメイン商品と多種多様なオプション商品によって成り立っているクラスタであると考えられる。

• Cluster-6

Cluster-6はMP2が極端に多く、続いてMP1、MP4が出現するクラスタである。これらのMPはいずれも片方向エッジから構成されるMPであり、商品の購入順が明確に決まっていることを示す。具体的には、定期的に発売される人気アーティストのCDや、コミックのシリーズ単行本などで観察されるパターンである。

• Cluster-7

Cluster-7はMP8及びMP7が極端に多いクラスタである。これらのMPはコアノードと周辺ノードの双方向エッジで構成されるMPであり、Cluster-2及びCluster-5のハイブリッドであるともいえる。つまり、1種類のメイン商品と多種多様な

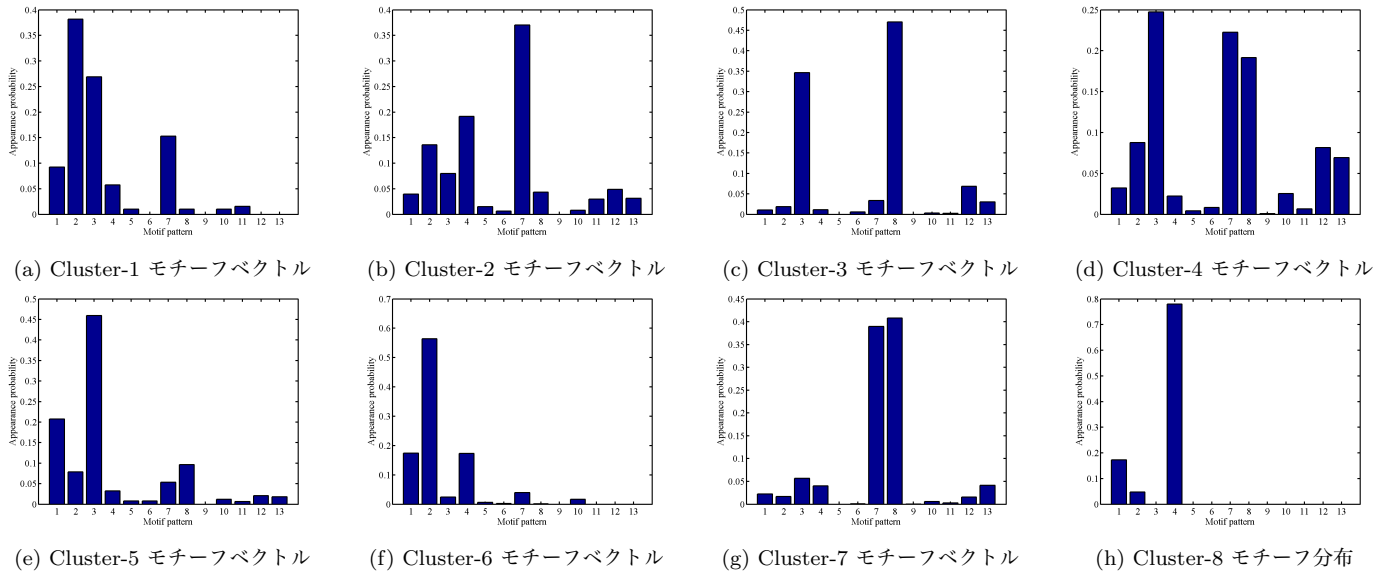


図 7 クラスタ別モチーフ分布 (Conventional-PHG)

オプション商品ないし多種多様なオプション商品と 1 種のオプション商品から構成されるパターンであると考えられる。加えて、コア商品と周辺商品が同時に購入された場合そのレビューの前後はユーザによって異なることから、双方向エッジは同時購入される商品にも観察されることが考えられる。

• Cluster-8

Cluster-8 は MP4, MP1, MP2 の片方向エッジ MP から構成される。この点で Cluster-6 と同様に購入の順序が明確に決まっている事が示されている。加えて極端に MP4 が多いことから 2 つの商品で購入が完結する傾向にあるといえる。これは小説の上下巻などで観察されるパターンである。

b) 主成分分析

構築した 141 モチーフベクトルに対し、主成分分析を行った。また、表 2 に示すように、累積寄与率が第四主成分で 85% に達していることから、分析の対象を第四主成分までとする。まず、第一主成分から第四主成分の固有ベクトルの各要素 (MP) について、正負それぞれ絶対値の大きいものを 2 つずつ抽出し、各主成分が示す特徴について分析する。固有ベクトルを図 8 に、各固有ベクトルから把握した構造的特徴を表 3 に示す。

• 第一主成分 (PC1)

大きな正の値を示したのは MP2 次いで MP4、負の値では MP8 次いで MP3 である。両者の違いは双方向エッジの有無である。MP2, 4 は単方向エッジのみから構成される MP, MP8, 3 は双方向エッジを含む MP である。さらに MP2, 4 のみに着目すると、MP2 は完全な順序性を示すのに対し、MP8 は順序性を一切示さない。以上のことから、第一主成分は順序性を表していると考えられる。

• 第二主成分 (PC2)

大きな正の値を示したのは MP3 次いで MP1、負の値では MP7 次いで MP4 である。エッジ方向について、MP3, 4 は 1 ノードから周辺 2 ノードへ流れ出る傾向にあり、MP7, 4 は周辺 2 ノードから 1 ノードへ集まる傾向にある。以上のことから、第

二主成分はコアノードと周辺ノードの間のエッジ方向を表していると考えられる。

• 第三主成分 (PC3)

大きな正の値を示したのは MP4 次いで MP1、負の値では MP2 次いで MP7 である。両者の違いは構造的対称性であると考えられる。MP4, 1 は線対称な MP であるが、MP2, 7 はいずれも非線対称な MP である。また、MP1 に僅差で続く MP8 も対称性を示す。以上のことから、第三主成分は MP の構造的対称性を表していると考えられる。

• 第四主成分 (PC4)

大きな正の値を示したのは MP7 次いで MP3、負の値では MP8 次いで MP2 である。これらの違いはコアノードから見た入エッジと出エッジの比であると考えられる。MP7, 3 において入出比はそれぞれ 2:1, 1:2 となっている。一方 MP8, 2 の入出比は 1:1 である。以上のことから、第四主成分はコアノードから見たエッジの入出比を表していると考えられる。

表 2 各主成分の寄与率及び累積寄与率 (Conventional-PHG)

	PC1	PC2	PC3	PC4
寄与率	0.3324	0.2655	0.1511	0.1022
累積寄与率	0.3324	0.5979	0.7490	0.8512

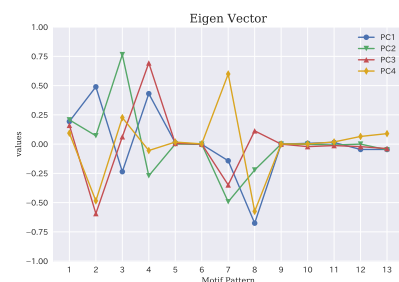


図 8 上位主成分の固有ベクトル (Conventional-PHG)

表 3 各主成分の特徴 (Conventional-PHG)

	正	負
PC1	順序性が高い	順序性が低い
PC2	放出型	流入型
PC3	対称性が高い	対称性が低い
PC4	入出比が大きい	入出比が小さい

表 4 クラスタ別 WCC 数 (Proposed-PHG)

Cluster-1	Cluster-2	Cluster-3	Cluster-4
42	43	27	14

4.3.3 Proposed-PHG の購買行動パターン分析

a) k -means クラスタリング

Proposed-PHG に対し、弱連結成分分解を行った。その結果、126WCC、7,744 ノード及び 6,126 エッジからなるグラフが構築された。この 126WCC からモチーフベクトルを取得し、 k -means クラスタリングを適用した。GAP 統計量 G_k により、クラスタ数は 4 とした。表 4 は、クラスタ別の WCC 数を示す。また、各クラスタのセントロイドのモチーフベクトルを図 9 に示す。

各クラスタの特徴について、MP の構造を軸に説明する。

• Cluster-1

モチーフベクトルに着目すると、MP2 が出現割合の殆どを占める。MP2 は完全な順序性を持ち、コアノードを経由する一方通行型である。

• Cluster-2

本クラスタでは、MP3 が最も多く出現し、続いて MP1 となる。これらの MP はいずれも中心となるコアノードから周辺ノードへと流出する傾向を持つ。このことから、スマートフォンとケースのような、1 つのメイン商品と複数のオプション商品の関係から成るクラスタであると考えられる。

• Cluster-3

Cluster-3 は MP7 が最も多く出現し、続いて MP3、MP8 となる。これらの MP はいずれも中心となるコアノードと周辺ノードとの関係から成り、また、MP7 は周辺ノードからコアノードへとエッジが流入してくる傾向にある。これは Cluster-2 と対象的な構造であり、複数種のメイン商品と、共通仕様のオプション商品の組み合わせであると考えられる。

• Cluster-4

Cluster-4 はその殆どが MP1 及び MP4 から成るクラスタである。これらの MP は単方向エッジかつ、コアノードと周辺ノードとの関係から成る MP である。このことから、Cluster-2 や 3 に比べ、順序性の強い商品に見られる購買行動パターンであると考えられる。

b) 主成分分析

構築した 126 モチーフベクトルに対し、主成分分析を行った。また、表 5 に示すように、累積寄与率が第四主成分で 85% に達していることから、分析の対象を第四主成分までとする。第一主成分から第四主成分の固有ベクトルの各要素 (MP) について、正負それぞれ絶対値の大きいものを 1 つないし 2 つずつ

抽出し、各主成分が示す特徴について分析する。固有ベクトルを図 10 に、各固有ベクトルから把握した構造的特徴を表 6 に示す。

• 第一主成分 (PC1)

大きな正の値を示したのは MP8 次いで MP3、負の値では MP2 次いで MP1 である。両者の違いは双方向エッジの有無である。MP2、1 は片方向エッジのみ、MP3,8 は双方向エッジを含む MP である。さらに MP2 及び 8 に着目すると、MP2 は完全な順序性を示すのに対し、MP8 は順序性を一切示さない。以上のことから、第一主成分は順序性の有無を表していると考えられる。

• 第二主成分 (PC2)

大きな正の値を示したのは MP3 次いで MP1、負の値では MP7 次いで MP8 である。まず、MP1、3 はコアノードから周辺 2 ノードへと流れ出る傾向にある。一方、MP7 は周辺 2 ノードからコアノードへと流入する傾向を持つ MP である。また、MP8 は完全な双方向エッジから成る MP であるが、さらに次点で続く MP4 は MP7 同様に周辺ノードからコアノードへの流入構造を持つ。このことから、第二主成分は、コアノードと周辺ノードの放出/流入の関係を表していると考えられる。

• 第三主成分 (PC3)

大きな正の値を示したのは MP1 及び MP4、負の値では MP2 次いで MP3 である。両者の違いは構造的対称性であると考えられる。MP1、4 は線対称な MP であるが、MP2、3 はいずれも非線対称な MP である。以上のことから、第三主成分は MP の構造的対称性を表していると考えられる。

• 第四主成分 (PC4)

大きな正の値を示したのは MP1 次いで MP7、負の値では MP4 次いで MP3 である。特に、MP1 及び MP4 に着目すると、いずれもコアノードと周辺ノードから成る MP であるが、エッジの向きが真逆である。このことから、第四主成分は順序性を伴うコアノードと周辺ノードとの関係を示していると考えられる。

表 5 各主成分の寄与率及び累積寄与率 (Proposed-PHG)

	PC1	PC2	PC3	PC4
寄与率	0.4002	0.2216	0.1692	0.0982
累積寄与率	0.4002	0.6218	0.7909	0.8891

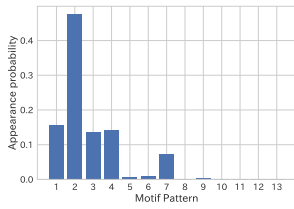
表 6 各主成分の特徴 (Proposed-PHG)

	正	負
PC1	順序性が低い	順序性が高い
PC2	放出型	流入型
PC3	対称性が高い	対称性が低い
PC4	片方向エッジかつ放出型	片方向エッジかつ流入型

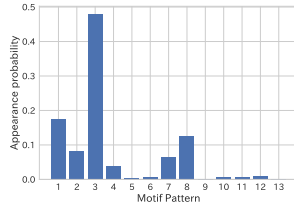
5. 考 察

5.1 購買行動パターンの比較・考察

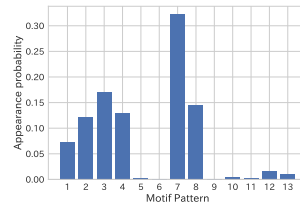
本節では、Conventional-PHG 及び Proposed-PHG から抽出した購買行動パターン及び主成分分析の結果を比較・分析す



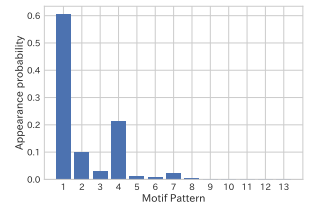
(a) Cluster-1 モチーフ分布



(b) Cluster-2 モチーフ分布



(c) Cluster-3 モチーフ分布



(d) Cluster-4 モチーフ分布

図 9 クラスタ別モチーフ分布 (Proposed-PHG)

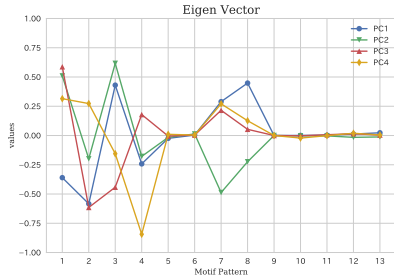


図 10 上位主成分の固有ベクトル (Proposed-PHG)

る。提案手法で構築した Proposed-PHG から抽出された 4 つの購買行動パターンは、従来手法で構築した Conventional-PHG から抽出された購買行動パターンと同様の性質を持つ。表 7 に、Proposed-PHG から抽出された購買行動パターンに対し、類似する Conventional-PHG の購買行動パターンを示す。また、主成分分析の結果から、「順序性」「(コアノードと周辺ノードからなる構造における) 放出・流入」の局所構造が、Conventional-PHG 及び Proposed-PHG 双方において重要であると示された。このことから、提案手法は従来手法と同程度に重要な構造を抽出可能であり、かつ、従来手法が分析対象とできないエッジを分析対象にすることが可能であると考えられる。

表 7 購買行動パターンの対応表

Proposed-PHG	Cluster-1	Cluster-2	Cluster-3	Cluster-4
Conventional-PHG	Cluster-6	Cluster-5	Cluster-2	Cluster-1

5.2 Z スコアの影響

5.1 で、提案手法は、従来手法では分析対象とできないエッジから、従来手法同様の知見を抽出可能であることを示した。これを踏まえ、本節では、出現頻度と Z スコアの 2 つの尺度を組み合わせた場合について考察する。頻度を 10^1 未満 (低頻度層)、 10^1 以上 10^2 未満 (中頻度層)、 10^2 以上 (高頻度層) の 3 層に分け、各層について、Z スコアの閾値を引き上げながら全体のモチーフベクトルの変化を観察する。これにより、各層について Z スコアの閾値別の特徴を把握したり、閾値 θ_z を適切に設定する材料とすることができる。

低頻度、中頻度、高頻度のモチーフベクトルの変化について、特に出現確率の高い 6 つの MP (MP1, MP2, MP3, MP4, MP7, MP8) を、図 11, 図 12 及び図 13 にプロットした。縦

軸が各 MP の出現確率、横軸が Z スコアの閾値、各線はそれぞれ一つの MP に対応している。

低頻度層のモチーフベクトルの変化を図 11 に示す。まず、Z スコアが 25 付近で、すべてのモチーフの出現確率が 0 になっている。これは、低頻度層のエッジのうち、 $Zscore(e(i,j)) > 25$ を満たすエッジ $e(i,j)$ が存在しなくなったことを意味する。また、興味深い点として、Z スコアの閾値が上がるに連れ、MP2 に比べ、MP1 及び MP4 が多く出現している。図 6 に示すように、MP1, 2, 4 はいずれも 3 本のエッジから構成される MP であり、エッジの追加や削除によって互に変化しないモチーフである。MP1, 4 はスマートフォンとスマートフォンケースのように一つのメイン商品と複数のオプション商品、ないしは複数のバリエーションをもつメイン商品と共通仕様のオプション商品から観察される構造である。つまり、MP1 及び MP4 は、周囲の商品との関係が強い商品関係を表しており、Z スコアの閾値 θ_z を引き上げることで、それらの商品関係が抽出可能になると考えられる。

高頻度層のモチーフベクトルの変化を図 13 に示す。ここでは、MP8 の推移に着目する。図 6 に示すように、MP8 は 4 本のエッジからなる複雑な構造である。Z スコアの閾値を上げることは、結果的にエッジの削除につながる。その為、低頻度層及び中頻度層では、Z スコアの閾値 θ_z の小さいうちから MP8 の出現確率は低くなっている。一方、高頻度層においては、Z スコアの閾値にかかわらず MP8 の出現確率は常に高い値を示している。これは、高頻度層に属する商品がそもそも非常に多くの商品と関係を持っていることで、 θ_z 上昇に伴うエッジの減少を経ても、なお MP8 の構造を維持できるためであると考えられる。つまり、 θ_z を上昇させた上で抽出できる MP8 構造は、大量の商品と多くの関係を持つメジャーな商品において、特に強い関係が築かれた商品関係であると考えられる。

6. おわりに

本研究では、グラフの局所構造に着目するグラフマイニング手法を提案した。具体的には、エッジの重みについて、隣接エッジ集合内で Z スコアを算出し、閾値以上の Z スコアを示したエッジを分析対象としてサブグラフを生成する。

楽天市場のレビューデータセットを購買順序とみなし、購買順序グラフを構築、従来手法である大域的な指標 (出現頻度) と、提案手法とした局所性に着目した指標 (Z スコア) それぞれについて分析対象を決定し、2 つのサブグラフを構築した。

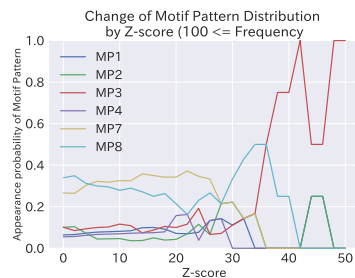
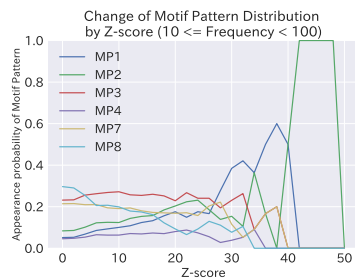
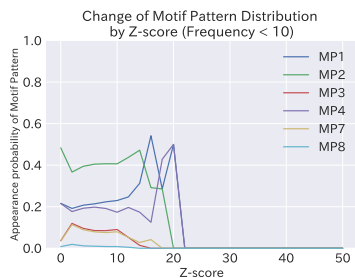


図 11 モチーフベクトルの変化 (低頻度層) 図 12 モチーフベクトルの変化 (中頻度層) 図 13 モチーフベクトルの変化 (高頻度層)

このサブグラフについて、モチーフ分析に基づき、典型的な購買行動パターンを抽出するグラフマイニングを行った。その結果、従来手法で生成したグラフでは 8 つの購買行動パターンが抽出された。加えて、それらの購買行動パターンを分類する観点として、「順序性」「(コアノードと周辺ノードから成る構造における) 放出・流入」「構造的対称性」「入出比」の 4 つの構造が重要であることが示された。また、提案手法でも、従来手法と同様の購買行動パターンが抽出され、「順序性」「コアノードと周辺ノードから成る構造における) 放出・流入」が重要な構造であることが示された。

これを踏まえ、従来手法と提案手法を組み合わせた分析として、出現頻度別に Z スコアを変化させた際のモチーフベクトルの変化の観察を行った。その結果、低頻度層及び高頻度層において、高い Z スコアを示すエッジは、コアノードと周辺ノードからなる構造を持つことが明らかになった。この構造は、購買行動パターン分析で重要な局所構造として抽出された構造である。以上のことから、出現頻度ベースの従来手法と Z スコアベースの提案手法を組み合わせることで、局所性に考慮しつつ、従来手法より分析対象を拡張し、重要な構造を抽出可能であることを示した。

今後の課題としては、購買順序グラフ以外のデータへの適用や、より適切な相対値の算出手法の考案などが挙げられる。

謝 辞

本研究は JSPS 科研費 JP16H02904 の助成を受けたものです。本研究の実装・評価に際し、楽天株式会社と国立情報学研究所が提供する「楽天データ」を利用しました。

文 献

- [1] P.-Y. Chen and A. O. Hero. Deep community detection. *IEEE Transactions on Signal Processing*, 63(21):5706–5719, 2015.
- [2] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.
- [3] A. Hayashi, M. Kohjima, T. Matsubayashi, and H. Sawada. Regularity measure and influence weight for analysis and visualization of consumer's attitude. In *2015 19th International Conference on Information Visualisation*, pages 290–299, July 2015.
- [4] F. Isinkaye, Y. Folajimi, and B. Ojokoh. Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16(3):261–273, 2015.
- [5] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda. Topic tracking model for analyzing consumer purchase behavior. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence, IJCAI'09*, pages 1427–1432, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.
- [6] J.-G. Liu, Q. Guo, and Y.-C. Zhang. Information filtering via weighted heat conduction algorithm. *Physica A: Statistical Mechanics and its Applications*, 390(12):2414–2420, 2011.
- [7] J.-G. Liu, T. Zhou, and Q. Guo. Information filtering via biased heat conduction. *Physical Review E*, 84:037101, 2011.
- [8] S. Salihoglu. Networks as vectors of their motif frequencies and 2-norm distance as a measure of similarity. Technical report, Stanford University, 2006.
- [9] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web, WWW '01*, pages 285–295, New York, NY, USA, 2001. ACM.
- [10] A. Srivastava. Motif analysis in the amazon product co-purchasing network. *Computing Research Repository*, abs/1012.4050, 2010.
- [11] P. Symeonidis, E. Tiakas, and Y. Manolopoulos. Product recommendation and rating prediction based on multimodal social networks. In *Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11*, pages 61–68, New York, NY, USA, 2011. ACM.
- [12] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.
- [13] F. Yu, A. Zeng, S. Gillard, and M. Medo. Network-based recommendation algorithms: A review. *Physica A: Statistical Mechanics and its Applications*, 452:192–208, 2016.
- [14] T. Zhou, Z. Kuscsik, J. Liu, M. Medo, J. Wakeling, and Y. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.
- [15] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang. Bipartite network projection and personal recommendation. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 76(4):046115+, 2007.
- [16] 稲福和史, 伏見卓恭, 佐藤哲司. レビュー順序グラフに基づく購買行動パターンの分析. 第 9 回データ工学と情報マネジメントに関するフォーラム, A3-4, 2017.