Vol.2016-DBS-163 No.18
Vol.2016-IFAT-123 No.18
2016/9/15

レビュー順序グラフを用いた購買行動パターンの抽出手法の提案

稲福 和史[1,a)]　伏見 卓恭[1,b)]　佐藤 哲司[1,c)]

概要：オンラインショッピングの拡大に伴い，膨大な顧客購買データが利用できるようになっている．本研究では，有向グラフで表したユーザの購買履歴をモチーフ分析することで，典型的な購買行動パターンを抽出する手法を提案する．提案手法では，購買履歴から Purchased History Graph(PHG) を構築し弱連結成分 (WCC) 分解を行う．得られた各 WCC 毎に算出したモチーフベクトルを特徴量として，k-means 法によりクラスタリングを行い購買行動パターンを抽出する．実運用されているオンラインショッピングのレビューデータを購買データとして提案手法を適用した結果，典型的な購買行動パターンを抽出できたため報告する．

キーワード：購買行動，モチーフ分析

Purchase Patterns Extraction Method Based on Motif Analysis of Directed Graphs

Kazufumi Inafuku[1,a)]　Takayasu Fushimi[1,b)]　Tetsuji Satoh[1,c)]

1. はじめに

オンラインショッピングの拡大に伴い，膨大な顧客購買データが利用できるようになっている．このデータを用いてユーザの購買行動の推定や商品推薦の研究が盛んに行なわれている [1], [2], [3], [4], [5]．オンラインショッピングを行うユーザは，商品を購入後にレビュー記事の投稿も日常的に行っている．このため，レビューデータの投稿順序はユーザの購買順序を表していると考えられる．本研究では，オンラインショッピングにおけるユーザの購買順序に着目し，典型的な購買行動パターンの抽出法を提案している．本稿の構成は次の通りである．提案手法の詳細な説明を第 2 章，実験の結果及び考察を第 3 章及び第 4 章，最後にまとめを第 5 章に示す．

[1] 筑波大学
[a)] s1411497@klis.tsukuba.ac.jp
[b)] fushimi@ce.slis.tsukuba.ac.jp
[c)] satoh@ce.slis.tsukuba.ac.jp

2. 提案手法

本研究では，レビューデータから構築するレビュー順序グラフ (PHG [*1]) のモチーフベクトルに基づく購買行動パターン抽出手法を提案する．ユーザー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 $R(u) = [r_1, r_2, \ldots, r_{N_u}]$ 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 $I(u) = [i_1, i_2, \ldots, i_{N_u}]$，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．

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

得られた SI を基に PHG を構築し分析する手順を以下に

[*1] レビューの投稿順序がユーザの購買順序を表しているとした事から，レビュー順序グラフを Purchased History Graph(PHG) と称する．

示す．

2.1 レビュー順序グラフの構築

本節では，各ユーザのレビューデータから有向グラフを構築する．ユーザ u が 3 種類 4 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 1** 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．

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

この $G = (V, E)$ で表される有向グラフを PHG [*2]とする．

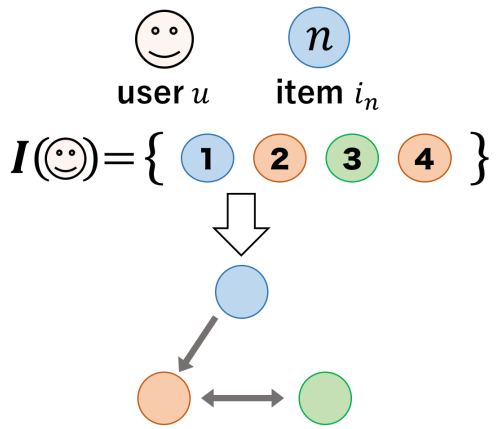


図 1 レビュー順序グラフ

2.2 弱連結成分分解

2.1 で構築した PHG に対し，弱連結成分分解を行う．PHG $G(V, E)$ 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．

2.3 モチーフベクトルの構築

本節では 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 2**)，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 MP 出現確率ベクトルを取得する．これをモチーフベクトル[*3]と称する．

[*2] 典型的な購買行動パターンの抽出するために，(i, j) ペアの出現頻度が 10 以上のものに対しエッジを追加している．
[*3] 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．

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「iPhone と iPhone 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．

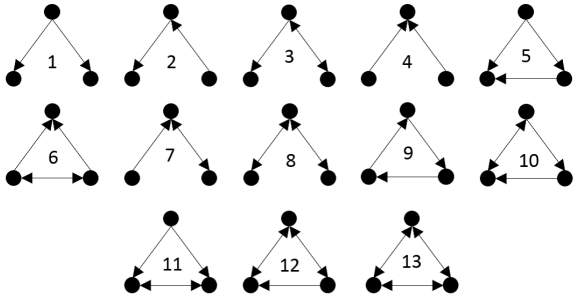


図 2 3 ノードモチーフパターン

2.4 k-means クラタリング

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

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

クラスタ数の決定は GAP 統計量 G_k[6] 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．

3. 実験

3.1 データセット

本研究では，Nii が研究用に公開している楽天市場のレビューデータ[*4]を使用した．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 レビューを評価データセットとした．

3.2 PHG の統計分析

3.1 のデータから構築した 10,266 ノード，13,939 エッジの PHG を**図 3** に示す．PHG は少数の大きなサイズの連

[*4] http://www.nii.ac.jp/dsc/idr/rakuten/rakuten.html

結成分と多数の小さなサイズの連結成分から構成されている．また，複数箇所に類似した構造が観察されることから，一定の構造パターンが存在すると考えられる．

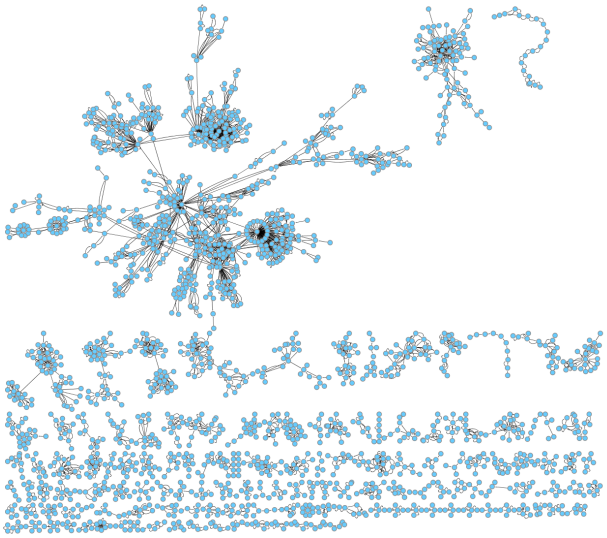


図 3 PHG の全体像

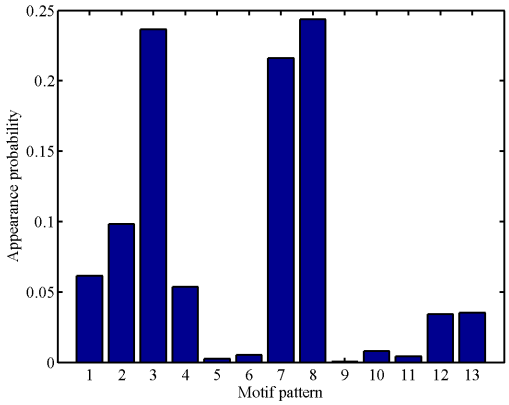


図 4 PHG モチーフベクトル

PHG 全体のモチーフベクトルは**図 4** に示すように双方向エッジを持つ MP3，MP7,MP8 が高い確率で出現している．双方向エッジを持つ 2 商品は購買の前後が明確ではなく，購買行動の順序関係を抽出する事が難しい．そこで，次のステップにおいて WCC 毎に詳細に分析する．

3.3 弱連結成分分解

3.2 で構築した PHG に対し弱連結成分分解を行った．この際，WCC のサイズが 5 以上のものを実験対象とした．結果，PHG は 141WCC に分解され，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．**図 5** は WCC のサイズ分布を示している．べき乗則の存在が確認できることから，少数の大規模な WCC と多数の小規模な WCC から成るグラフだといえる．

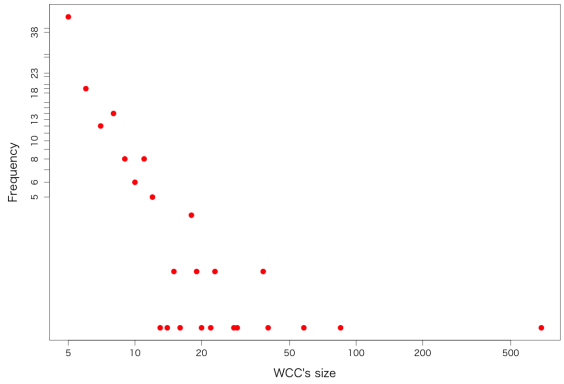


図 5 WCC サイズ分布

3.4 モチーフ分析及び k-means クラスタリング

3.3 で構築した WCC 群に対し，それぞれモチーフ分析を行い 141 個のモチーフベクトルを取得した．これに対し，PHG と同様に正規化 ($L_1 norm$) し，k-means クラスタリングを行った．GAP 統計量 G_k によりクラスタ数 K は 8 とした．表 1 は，クラスタ別の WCC 数を示す．また，**図 8** は，各クラスタのセントロイドベクトルを示す．

表 1 クラスタ別 WCC 数

Cluster-1	Cluster-2	Cluster-3	Cluster-4
11	14	20	34
Cluster-5	Cluster-6	Cluster-7	Cluster-8
26	10	20	6

- Cluster-1
  モチーフ分布に着目すると，もっとも多く出現するモチーフパターンが 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，**図 6** のように特定の商品 (コアノード) を中心に購買行動を次々と発展させているといえる．これを踏まえると，Cluster-1 は「中心経由発展型」の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，第二位に MP2，第三位に 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集中型であると考えられる．具体的には，**図 7** のように多種多様なメイン商品群とそれらに共通して使用出来る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

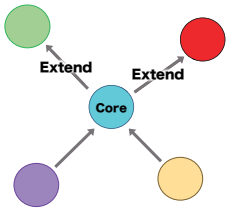


図 6 中心経由発展型 模式図

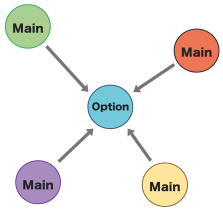


図 7 コアノード集中型 模式図

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恐れがあった．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 も中心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，短いシークエ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と考えられる．
- 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．

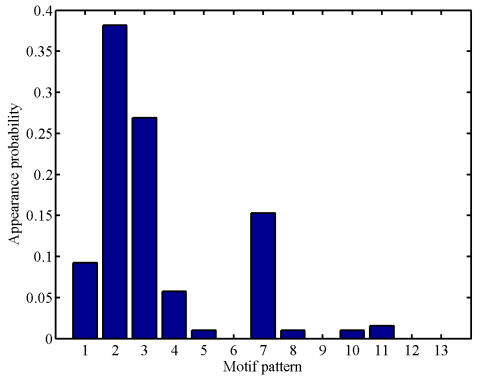
4. 考察

4.1 クラスタの複雑度

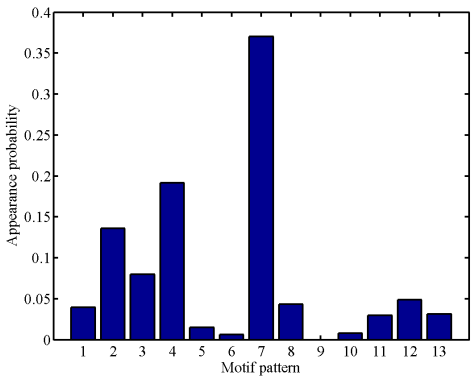
最も基本的な購買行動パターンは 1 つの MP によって表されることが望ましい．しかし実際に観察される購買行動パターンは複数の MP の混合からなる．ここで実パターンの複雑度をモチーフベクトルの各要素の累積によって定量化する．モチーフベクトルは確率ベクトルであるため，全要素を累積すると 1 になる．要素の値が大きい順に累積していき 1 に達するまでの速さを $RWCC_k$ 毎に比べたところ，**図 10** のようになった．急速に 1 に近づく RWCC は少数の MP から構成されているといえる．図 10 によれば，Cluster-8 が最も早く 1 に到達している．この結果は，図 8 とも一致しており，Cluster-8 は非常にシンプルなパターンであるといえる．一方，cluster2 や cluster4 は 1 に到達するのに多くの MP を要している．こちらも図 8 の結果と一致しており，Cluster-2,4 は複雑なパターンであるといえる．これらの結果に基づき，購買行動パターンを抽出する際の閾値を設定する．例えば，閾値を MP3 種類で累積確率が 0.8 以上と設定した際，Cluster-8,3,6,7 が実購買行動パターンとして抽出される．閾値を状況によって変更することにより，実購買行動パターンの抽出が可能になる．

4.2 実際の購買行動パターン

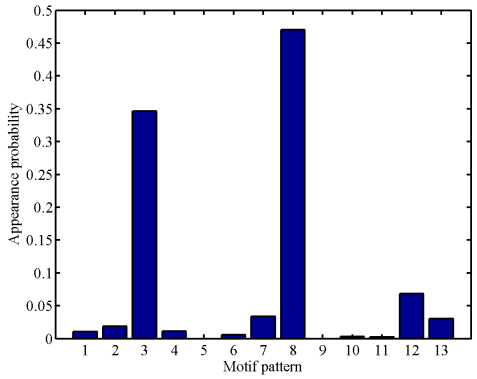
ここでは，PHG から実際に観察された購買行動パターン (APP:Actual Purchased Pattern) を 3 つ紹介する．APP-1 は人気ミュージシャンの CD のグラフである．この購買順序は CD の発売日と一致している．つまり，一定数のファンが CD が発売されるたびに新作を購入した結果であり，3.4 における cluster-6 に属すると考えられる．APP-2 はランタンと電池のグラフである．コアノードが電池，周辺ノードがランタンである．ランタンは，明るさの違いやサイズの違いなど様々なバリエーションが存在する．これは，様々な種類のメイン商品群と共通して使用出来るオプション商品の組み合わせであり，cluster-2 に属すると考え



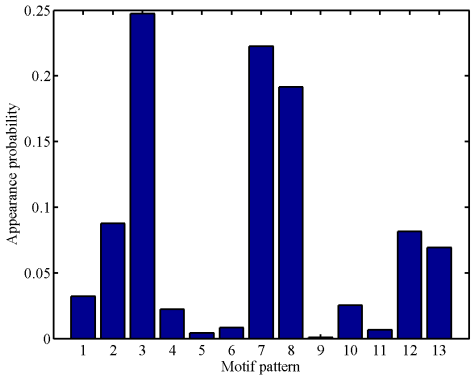
Cluster-1 モチーフ分布



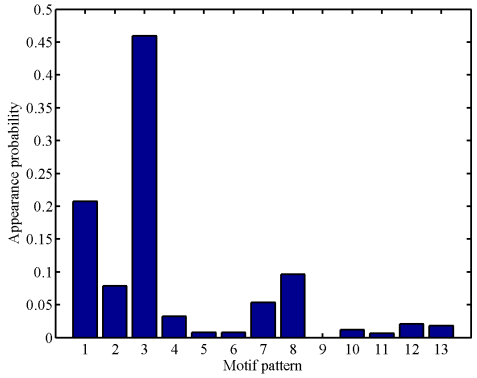
Cluster-2



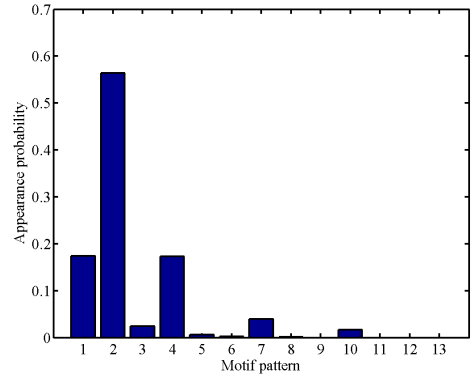
Cluster-3



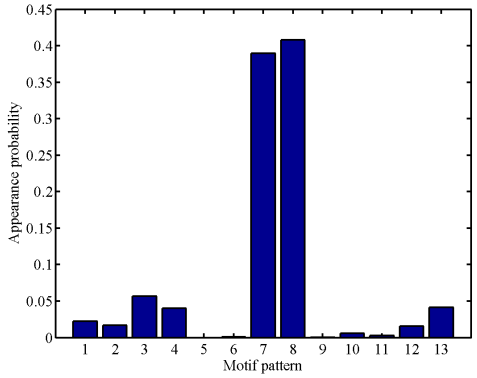
Cluster-4



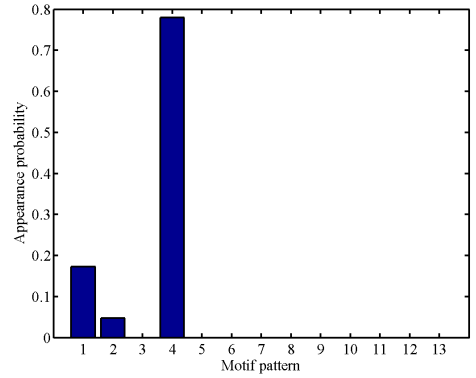
Cluster-5



Cluster-6



Cluster-7



Cluster-8

図 8 クラスタ別モチーフ分布

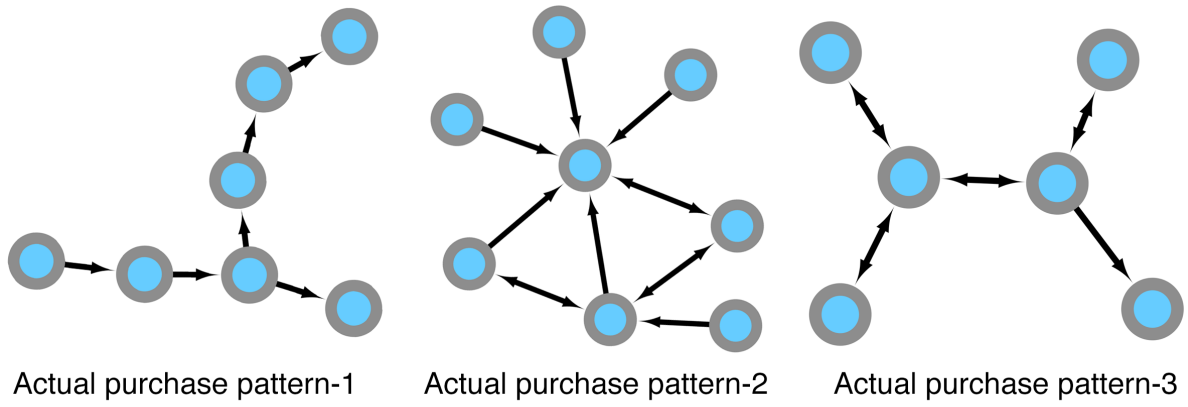


図 9 Actual purchase pattern

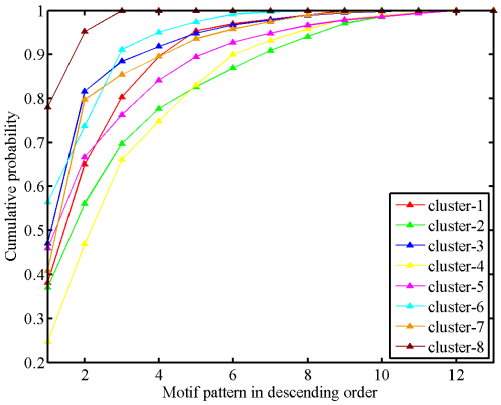


図 10 複雑度

られる．APP-3 はダニ対策布団用品のグラフである．中央の 2 ノードはいずれも掛け布団カバーであり，両者の違いはそのサイズである．周囲ノードはベッドシーツ及び敷布団カバーである．こちらも左右両者の違いはサイズである．中心ノードと周囲ノードの間には双方向エッジが観察される．これらの商品は同時購入される傾向にあり，ユーザによってレビューの前後が入れ替わることによって双方向エッジが観察できると考えられる．これは，Cluster-7 に属すると考えられる．他にも，同様のグラフ構造を持つが商品自体は全く異なるジャンルのものであるなど典型的な購買行動パターンの存在が確認出来た．

5. まとめ

本研究では，レビューデータを用いたモチーフ分析及び k-means クラスタリングによる購買行動パターンの抽出法を提案した．楽天市場のレビュー履歴から有向グラフを構築し，WCC 毎にモチーフ分析，取得したモチーフベクトルに k-means クラスタリングを行う事で 8 つの購買行動パターンを抽出した．さらに，購買行動パターンの複雑度を算出する事でよりシンプルな購買行動パターンを抽出できた．研究課題としては，レビューのタイムラグについての検討，様々なデータセットでの検討，閾値の設定に関する検討が残った．

謝辞

本研究は，JSPS 科研費 16H02904 の助成を受けたものです．また，楽天株式会社が国立情報学研究所の協力により研究目的で提供している「楽天公開データ」を利用しました．ここに記して謝意を示します．

参考文献

[1] 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.

[2] F. Isinkaye, Y. Folajimi, and B. Ojokoh. Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16(3):261 – 273, 2015.

[3] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda. Topic tracking model for analyzing consumer purchase behavior. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, pages 1427–1432, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[4] 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.

[5] P. Symeonidis, E. Tiakas, and Y. Manolopoulos. Product recommendation and rating prediction based on multi-modal social networks. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 61–68, New York, NY, USA, 2011. ACM.

[6] 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.