2L4-2

# グラフ構造による相関ルールの視覚化ツール:KIZUNA

Visualization tool for association rules in graph structure: KIZUNA

岩﨑幸子 *1 Sachiko Iwasaki　中元政一 *2 Masakazu Nakamoto　中原 孝信 *3 Takanobu Nakahara　宇野 毅明 *4 Takeaki Uno　羽室 行信 *5 Yukinobu Hamuro

*1*2*5関西学院大学 経営戦略研究科
Institute of Business and Accounting, Kwansei Gakuin University

*3専修大学 商学部
School of Commerce, Senshu University

*4国立情報学研究所 情報学プリンシプル研究系
Principles of Informatics Research Division, National Institute of Informatics

Association rule mining is actively carried out in variouse business fields, especially it is popular as a market busket analysis in retail industry. However it often results in tons of rules found, and it forces analysts to do a lot of cumbersome tasks. In this paper we introduce a new methodology to choose the important association rules and visualize them in a form of graph structure.The main feature of this methodology is that it selects association rules by mutual-ranking measure between two items, as a result, we succeeded to visualize both of important individual rules and global relation of the rules. We also impremented it as a web application software named "KIZUNA". We demonstrate some cases in applying the tool to a large scale purchasing data.

## 1. はじめに

1990 年代に示された「おむつ」と「ビール」の相関ルールの事例は、意外性のあるルールの発見とクロスマーチャンダイジングという店頭施策の打ち出しやすさから、探索的な知識発見の手法であるデータマイニングを一躍有名にさせた。それから 20 年以上が経過した現在でも、流通・小売業ではデータ利活用の 1 つとして, 相関ルールを用いた分析が盛んに取り組まれており、その結果を利用したクロスセルや棚割りの決定などへの応用が期待されている。しかしながら、列挙されるルールが膨大になったり、全体を俯瞰できる視覚化方法がないなどの問題がある。そこで我々は、これらの問題を解決するための視覚化手法を開発し、web アプリケーション「KIZUNA」として実装した *1。本稿ではその手法と応用ケースについて論じる。

## 2. 視覚化の評価基準

相関ルール分析は、データマイニングの分野で代表的な分析手法で、特にルールを高速に列挙する技術は飛躍的な進展を遂げてきた。しかしながら、パラメータの設定次第では時に大量のルールが出力され、そこから興味深いルールを抽出するまでにユーザに多大な負担を強いることも少なくない。このような問題を解決すべく、相関ルールの視覚化に関する研究が注目されてきた。代表的な視覚化としては、評価指標を 2 軸にとった散布図による方法、条件部と結論部を行と列に示した行列形式による方法、アイテム集合を複数のアイテム軸とアイテム間のリンクで表現する方法、アイテムを節点、辺をルールの強さとしたグラフ構造による方法、などがあげられる [3],[4]。それぞれの手法には一長一短あり、目的に応じて使い分けることが現実的な対応であろう。しかしながら、多くの視覚化に共通した問題点として、個別ルールの把握 (ローカル性) とルールの全体的な関係性の俯瞰 (グローバル性) を同時に実現することが困難であることが挙げられる [2]。ローカル性とグローバル性の両立という観点からは、グラフ構造による視覚化が優れており、本研究でもグラフ構造による視覚化を採用している。しかしながら、ルール数が増えてくると、辺が錯綜して見にくくなったり、特定のアイテムへの接続が集中したりするといった欠点がでてくる。本研究ではグラフ構造の利点は享受しながら、このような欠点を克服するルールの選択方法を提案するものである。

以上の議論を踏まえ、本研究では、以下に示す網羅性、連結性、分散性の 3 つの評価軸を「相関ルール視覚化評価基準」として設定することにする。

**網羅性** できるだけ多くのアイテムを含む相関ルールが選択されている。

**連結性** アイテムをできるだけ多く連結し全体的な関係性を俯瞰できる。

**分散性** 特定のアイテムに接続が極度に集中しない。

## 3. 手法

アイテム集合 $I$ についてのトランザクションデータベース $D = \{T_1, T_2, \cdots, T_n | T_i \subseteq I\}$ から得られる相関ルールは、アイテム集合 $X, Y \subseteq I$ について $X \Rightarrow Y$ と表現される。小売店のケースでは、アイテムを商品に、トランザクションをレシートに対応させれば考えやすい。提案手法は、$D$ から列挙される多数の相関ルールから少数の有用なルールを選択し、視覚化するものであり、以下に示す 3 つのフェーズから構成される。

**1) 相関ルール列挙** まず最初に、表 1 に示されるような相関ルールの評価指標についての下限値を与えることで、その条件を満たす相関ルール集合 $R$ を作成する。またルールの解釈の容易性を優先するため、条件部と結論部のアイテム集合のサイズを 1 に限定する ($|X| = |Y| = 1$) *2。

**2) ランクによるルール選択** フェーズ 1 は一般的な相関ルールの列挙手法であり、与える評価指標の下限値によっては多数

連絡先: 宇野毅明, 国立情報学研究所 情報学プリンシプル研究系, 東京都千代田区一ツ橋 2-1-2, TEL:03-4212-2000, uno@nii.jp

*1 http://www.nysol.jp で公開している。

*2 実装系では、2 アイテムに限定した頻出アイテム集合を高速に列挙する SSPC(Similar Set Pair Comparison) を利用している [5]。

表 1: KIZUNA で利用可能な評価指標一覧

| 評価指標 $sim$ | 関数表記 | 内容 | 定義 | 値域 |
|---|---|---|---|---|
| supp | $\mathrm{supp}(X \Rightarrow Y)$ | 支持度 | $\mathrm{p}(X, Y)$ | 0..1 |
| conf | $\mathrm{conf}(X \Rightarrow Y)$ | 確信度 | $\frac{\mathrm{p}(X,Y)}{\mathrm{p}(X)}$ | 0..1 |
| npmi | $\mathrm{npmi}(X \Rightarrow Y)$ | 標準化 PMI | $\frac{\log(\mathrm{p}(X)\mathrm{p}(Y))}{\log(\mathrm{p}(X,Y))} - 1$ | −1..1 |
| jacc | $\mathrm{jacc}(X \Rightarrow Y)$ | Jaccard 指数 | $\frac{\lvert\mathrm{occ}(X)\cap\mathrm{occ}(Y)\rvert}{\lvert\mathrm{occ}(X)\cup\mathrm{occ}(Y)\rvert}$ | 0..1 |

$\mathrm{occ}(X) = \{T_i | X \subseteq T_i\}$

$\mathrm{p}(X) = |\mathrm{occ}(X)|/|D|$, $\mathrm{p}(X, Y) = |\mathrm{occ}(X) \cap \mathrm{occ}(Y)|/|D|$

のルールが列挙される。そこで次に、$R$ から、同じく表 1 に示される評価指標のランク情報を用いて部分ルール集合 $R'$ を選択する。その方法が本提案手法の核心である。選択方法は次の通りである。まず $R$ の任意の相関ルール $X \Rightarrow Y$ の評価指標 $sim$ によるランク $\mathrm{rank}_{sim}(X \Rightarrow Y)$ を式 (1) の通り定義する。これは、条件部に $X$ を持つ部分ルール集合での、ルール $X \Rightarrow Y$ の評価順位を表している。

$$\mathrm{rank}_{sim}(X \Rightarrow Y) = |\{i : sim(X \Rightarrow Y) \leq sim(X \Rightarrow \{i\}), \; sim(X \Rightarrow \{i\}) \in R\}_{i \in I}| \quad (1)$$

ランクによるルール選択の方法は 2 つあり、一つは、「両想い (friend) 相関ルール」と呼称するルールで、式 (2) の条件を満たすような相関ルール $X \Rightarrow Y$ である。このようなルールは、$X$ と $Y$ が相互に強い関連性を持っていることを意味する。

$$\mathrm{rank}_{sim}(X \Rightarrow Y) \leq k \text{ and } \mathrm{rank}_{sim}(Y \Rightarrow X) \leq k \quad (2)$$

もう一つの選択方法は、「片想い (pal) 相関ルール」と呼称するルールで、式 (3) の条件を満たすようなルールであり、$X$ から見れば $Y$ は強い関連を持つが逆はそうでないようなルールである。

$$\mathrm{rank}_{sim}(X \Rightarrow Y) \leq k \text{ and } \mathrm{rank}_{sim}(Y \Rightarrow X) > k \quad (3)$$

**3) グラフ構造による視覚化** 最終的に選択された相関ルール集合 $R'$ を有向グラフ $G = (V, E)$ で表現する。頂点集合 $V$ は $R'$ の相関ルールの条件部と結論部の和集合で $V = \bigcup_{r \in R'} \{X, Y | X$ は相関ルール $r$ の条件部, $Y$ は結論部 $\}$ である。また辺集合 $E$ は相関ルールの (条件部, 結論部) を有向辺とした集合で、$E = \bigcup_{r \in R'} \{(X, Y) | X$ は相関ルール $r$ の条件部, $Y$ は結論部 $\}$ である。また、その定義から $X \Rightarrow Y$ が両想いルールであれば、$Y \Rightarrow X$ も両想いルールであるので、頂点 $X$ と $Y$ に $(X, Y)$ と $(Y, X)$ の 2 つの有向辺を持つことになるが、視覚化する際には、両方向の矢印を持った一つの辺で表現する。以上の方法によって得られたグラフを「相関ルールグラフ」と呼ぶ。

# 4. 実験

## 4.1 実験方法

本手法を用いた視覚化で設定が必要なパラメータは、評価指標 ($sim$) の選択、ランク ($k$)、そして両想い/片想いの選択、の 3 つである。そこで、これらの値によってどのようなグラフが生成されるかについてのいくつかの実験結果を示す。ただし、ここでは片想いルールについては扱わず、両想いルールのみを対象とすることにする。また、評価指標の定義から、$\mathrm{rank}_{\mathrm{supp}}(X \Rightarrow Y) = \mathrm{rank}_{\mathrm{conf}}(X \Rightarrow Y)$ であるので、支持度と確信度による両想い/片想い相関ルールは同じになる。そのため、確信度による実験は省いている。実験で用いるデータは表 2 に示される (株) マクロミルより提供された消費者購買履歴データである。

表 2: QPR データセット

| 項目 | 内容 |
|---|---|
| 対象店舗 | 日本全国のスーパー、コンビニなど全 637 小売企業 |
| データ期間 | 2012 年 1 月 1 日〜2014 年 6 月 30 日 |
| アイテム | JICFS 細分類コード ($\lvert I\rvert = 951$) |
| トランザクション | レシート ($\lvert D\rvert =$ 約 920 万) |

実験は、提案手法と一般的な相関ルール分析の手法とを比較することで実施する。一般的に行われる相関ルール分析では、各種評価指標を閾値として与えることでルールの絞り込みを行う (以下「一般手法」と呼称)。これは、提案手法のフェーズ 1 のみを実施することに相当する。

そこで、提案手法と一般手法の比較を次のように行うことにする。まず、両手法ともフェーズ 1 において、最小支持度に同じ下限値を用いてルール集合 $R$ を取得する。提案手法では、フェーズ 2 において評価指標 $sim$、ランク $k$ の両想い相関ルール集合 $R'$ を求める。一方で一般手法では、評価指標 $sim$ について $\mathrm{rank}(X \Rightarrow Y) \leq |R'|$ を満たすようなルール集合 $R''$ を $R$ より選択する。すなわち、提案手法で得られたルールと同じ数のルールを、同じ評価指標の上位から選ぶということである。なお、$R''$ は片思い相関ルールとは異なり、式 (3) の左項 $\mathrm{rank}(X \Rightarrow Y) \leq k$ のみでルール選択したことに相当する。このようにして得られた両者の相関ルール集合 $R', R''$ を、フェーズ 3 の方法で相関ルールグラフに変換する。

次に、構築された相関ルールグラフの評価方法を示す。グラフの良さは先に示した、網羅性、連結性、分散性の 3 点によって評価するが、それぞれ、アイテムカバー率、連結成分数、最大次数の 3 つのグラフ特徴量を評価値として用いる。

アイテムカバー率とは、グラフがどれだけ多くのアイテムをカバーしているかを示す指標で、ルール集合 $R, R', R''$ から作られる相関ルールグラフの節点集合をそれぞれ $V, V', V''$ とすると、提案手法におけるアイテムカバー率は $|V'|/|V|$(一般手法は $|V''|/|V|$) で定義される。この値が大きいほど、網羅性が高く優れた視覚化であると評価できる。

連結成分数は、得られた相関ルールグラフにおいて、任意の 2 点間にパスが存在する極大部分グラフの数であり、アイテムカバー率が同じであれば、その数が少ない方が全体の関係性をより俯瞰できていることになり、連結性が高いと評価できる。

そして最大次数は、相関ルールグラフ上の全アイテムの中で最も大きな次数 (edge の数) である。この値が大きいということは、特定のアイテムに関するルールが集中的に列挙されていることになる。他の評価項目が同じであれば、最大次数がより小さいほうが分散性が高いと評価する。

## 4.2 実験結果

$k = 1, 2, 3, 4, 5, 10, 20, 50$ とし、評価指標 $sim$=supp, jacc, npmi の三つについて比較した結果を表 3 に示す。いずれのパラメータにおいても、アイテムカバー率は、一般手法に比べて 2〜3 倍程度高くなっており、提案手法が、より全体を網羅した視覚化ができていることが分かる。また最大次数は、提案手法においては大きくても $k$ に抑えられるため、一般手法に比べて小さい値となっている。一方で連結成分数についてはパラメータ依存である。$sim$=supp の場合と、$k$ の値が小さい場合に、一般手法の方が小さくなる。一般手法において支持度でルールを選ぶと、単一アイテムとしての支持度が大きいアイテムに接続されるアイテムが多くなるためであろう。また $k$ が小さいと、そもそもルール数が少ないので、いずれの方法においても可読性の高いグラフが生成される可能性が高い。ただし、一般手法の場合、$k$ の値によっては最大次数が高くなり、辺が

錯綜したグラフが出力され可読性が低くなる可能性が高い。

表 3: 提案手法 (rank) と一般手法 (th) の比較。一行目で例示すると次の通り。評価指標 $sim$=supp、ランク $k=1$ の条件で両想いルールを選択すると 7 つのルールが得られ (edge 数)、その相関ルールグラフの評価値は、アイテムカバー率=0.045、最大次数=1、連結成分数=7 であった。一方で支持度 (supp) の大きい順に 7 つのルールを選択すると、それらの評価値が、0.026、7、1 であった。$R$ に含まれるアイテム数 $|V|=307$。

| $sim$ | $k$ | edge 数 | アイテムカバー率 | | 最大次数 | | 連結成分 | |
|---|---|---|---|---|---|---|---|---|
| | | $\|R'\|, \|R''\|$ | rank | th | rank | th | rank | th |
| supp | 1 | 7 | 0.045 | 0.026 | 1 | 7 | 7 | 1 |
| | 2 | 13 | 0.078 | 0.029 | 2 | 8 | 11 | 1 |
| | 3 | 22 | 0.104 | 0.042 | 3 | 12 | 11 | 1 |
| | 4 | 30 | 0.123 | 0.055 | 4 | 15 | 13 | 1 |
| | 5 | 41 | 0.149 | 0.068 | 5 | 19 | 14 | 1 |
| | 10 | 122 | 0.263 | 0.127 | 10 | 38 | 18 | 1 |
| | 20 | 397 | 0.456 | 0.257 | 20 | 74 | 18 | 1 |
| | 50 | 2096 | 0.811 | 0.586 | 50 | 176 | 9 | 1 |
| jacc | 1 | 52 | 0.338 | 0.127 | 1 | 10 | 52 | 12 |
| | 2 | 110 | 0.524 | 0.179 | 2 | 17 | 56 | 10 |
| | 3 | 153 | 0.589 | 0.221 | 3 | 22 | 48 | 14 |
| | 4 | 188 | 0.641 | 0.254 | 4 | 23 | 42 | 13 |
| | 5 | 236 | 0.713 | 0.296 | 5 | 26 | 33 | 18 |
| | 10 | 426 | 0.771 | 0.358 | 10 | 34 | 12 | 19 |
| | 20 | 901 | 0.820 | 0.589 | 20 | 50 | 7 | 16 |
| | 50 | 2860 | 0.912 | 0.840 | 50 | 88 | 2 | 7 |
| npmi | 1 | 80 | 0.521 | 0.328 | 1 | 6 | 80 | 31 |
| | 2 | 167 | 0.726 | 0.465 | 2 | 11 | 69 | 30 |
| | 3 | 252 | 0.840 | 0.534 | 3 | 17 | 51 | 31 |
| | 4 | 319 | 0.863 | 0.560 | 4 | 21 | 25 | 30 |
| | 5 | 396 | 0.876 | 0.625 | 5 | 26 | 15 | 29 |
| | 10 | 763 | 0.938 | 0.745 | 10 | 40 | 5 | 14 |
| | 20 | 1472 | 0.986 | 0.892 | 20 | 66 | 3 | 12 |
| | 50 | 3623 | 1.000 | 1.000 | 50 | 97 | 2 | 3 |

次に、グラフ構造の外観をパラメータ別により比較する。図 1 に、$sim$=npmi で $k=1,5,10,20$ の両想い相関ルールグラフを示す。いずれも同じレイアウトアルゴリズムで描画している。$k=5,10,20$ で網羅性、連結性、分散性の全てが比較的高いことが視覚的にも確認できるであろう。$k=1$ では連結性が低く、単体のルールが多く列挙されていることが視認できる。

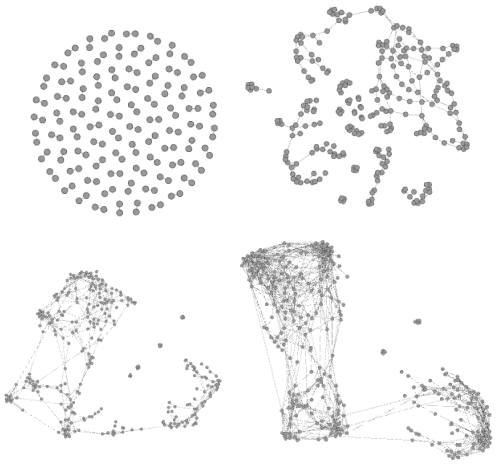

図 1: 評価指標 $sim$=npmi におけるランク $k$ による両想い相関ルールグラフ。左上 ($k$=1), 右上 ($k$=5), 左下 ($k$=10), 右下 ($k$=20)。グラフ描画ソフト Gephi で描画している [1]。

図 2 は、$sim$=supp,jacc で $k=10$ の両想い相関ルールグラフを示したものである。図 1 の左下のグラフも合わせて考察すると、npmi と jacc は比較的連結性が高く網羅的である一方、supp は網羅性が低く、連結性が低いことが視覚的に確認できる。

最後に、図 3 は、$k=10$ に対応する一般手法で視覚化したものである。提案手法に比べ、$sim$=npmi,jacc においては、いずれの相関ルール視覚化評価基準においても劣っていることが視覚的にも確認できよう。一方で $sim$=supp については、連結性を優先すれば閾値の方法が優れ、その他の点においては提案手法の方が優れていることが確認できる。

図 2: ランク $k$=10 を固定し場合の $sim$ =supp と $sim$ =jacc による両想い相関ルールグラフ。左 ($sim$=supp), 右 ($sim$=jacc)。

図 3: 一般手法での相関ルールグラフ。左から $sim$=supp, jacc, npmi。提案手法の $k$=10 に対応した辺数 (それぞれ、122, 426, 763) に合わせて各評価指標上位のルールを選択したもの。

## 5. 応用ケース

表 2 に示される QPR データを使用し、KIZUNA による視覚化を行い、その分析を実施した。最小支持度=0.001%、最小標準化 PMI=0.1 の条件で, $sim$=jacc、$k=5$ の両想いルールと $sim$=npmi、$k=1$ の片想いルールを同時に相関ルールグラフとして描画した (図 4)。$|V|=382$、$|E|=460$ で、27 個の連結成分数を得た。各連結成分は同質性が高く同時購買の意味解釈が可能なものが多い。全体として、「食品系」、「住居関連系」、そして「医薬品系」の 3 つのグループに識別できた。このように、全体を俯瞰して意味解釈が容易な視覚化を可能とするのが KIZUNA の主要な特徴である。

次に、ビジネスへの応用場面として、小売業において消費者が買い回りしやすい売場配置や関連販売 (クロスセル) を実施する場面を想定し、知見を得るために考察する。

図 5 は、図 4 における菓子と加工食品の相関ルールを拡大表示した図である。本結果をみると、スナック菓子とインスタントのカップ麺が共に共起性が高く双方向の関係を示している。同時購買の背景として、おやつなどの間食のニーズが考えられる。菓子売場と加工食品売場で販売場所は離れていても一緒に買われている。間食ニーズに応えるための売場配置の工夫やクロスセルなどで双方の購買促進をはかる施策などが本結果から検討できる。

図 6 は、図 4 における冷凍食品と加工食肉の相関ルールを拡大表示した図である。冷凍調理品に対し畜肉ソーセージが単方向の関係を示しており、同じ食肉売場の畜肉ハムとベーコンではなく冷凍食品への関係を持っている。その背景として考えられるのは、冷凍調理品はお弁当用の商材が多く、お弁当つながりでの関係を持っていることが考えられる。冷凍食品売場と

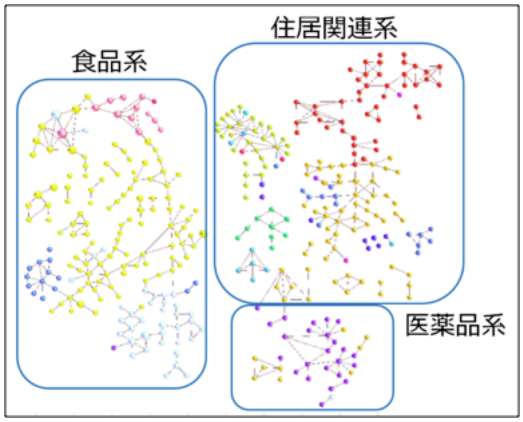


図 4: KIZUNA から得たグラフの全体図。枠線はわかりやすさのため事後的に加筆したものである。

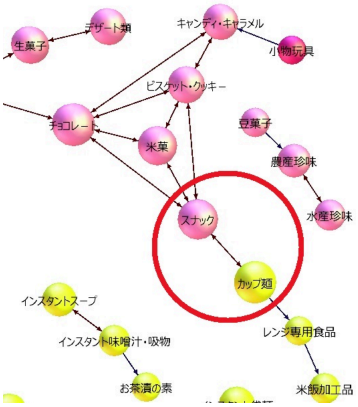


図 5: 菓子と加工食品の相関ルール (図 4 の拡大図)。

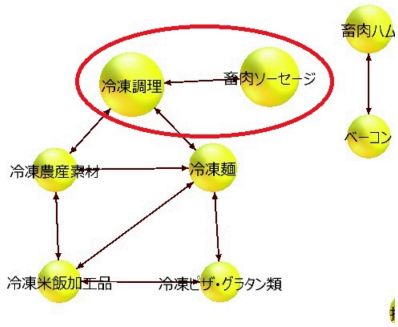


図 6: 冷凍食品と加工食肉の相関ルール (図 4 の拡大図)

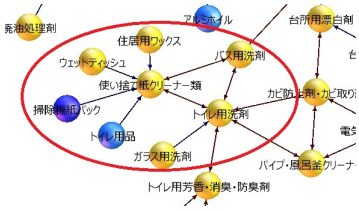


図 7: 住居清掃用品の相関ルール (図 4 の拡大図)

食肉の加工食肉売場は離れていることが多いが、本結果から、畜肉ソーセージならびに冷凍調理品の購買者に対し「お弁当」をテーマとした推奨施策を講じるなどのアイデアを得られる。

図 7 は、図 4 における住居清掃用品の相関ルールを拡大表示した図である。使い捨て紙クリーナーの部分に注目すると、使い捨て紙クリーナーはバス用洗剤とトイレ用洗剤と双方向の関係を示し、トイレ用品や掃除機紙パックから単方向の関係を示している。使い捨て紙クリーナーは、所謂日用品部門の住居用洗剤売場に配置されていることが多く、トイレ用品や掃除機の紙パックは雑貨部門の清掃関連の売場に配置されていることが多い。商物流も異なるため、担当バイヤーも異なる場合が多く、異なる売場に配置されることが一般的である。しかしながら、この結果から読み取れることは、消費者にとっては「掃除/清掃」に関わる製品群であるということで、これらは同類視されている可能性がある。消耗性の高い紙クリーナーと掃除機紙パックの売場配置を再検討し、トイレ用品の雑貨売り場には、使い捨て紙クリーナーのダブル配置を行うなどの配慮を講じるなど、売場改善策を本結果から考察できる。

## 6. おわりに

以上、ランク情報に基づいた相関ルールの視覚化手法について論じ、その有効性を、網羅性、連結性、分散性によって評価した。大規模な実購買履歴データを使って実験し、一般的な閾値により選択する手法に比べて、優れた視覚化性能を示すことを明らかにした。また視覚化されたグラフについて、その意味解釈をおこない、クラスタ間のつなぎの役割を担っているアイテムの発見が容易になるなど、提案手法による視覚化の有効性を示した。

## 謝辞

(株) マクロミルより消費者購買履歴データ QPR を提供いただいた。また中央大学の生田目崇教授には多大なご協力をいただいた。ここに感謝の意を表する。本研究の一部は、JST CREST の研究助成を受けている。

## 参考文献

[1] Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

[2] Couturier, O., Hamrouni, T., Yahia, S. B., and Nguifo, E. M. (2007, July). A scalable association rule visualization towards displaying large amounts of knowledge. In IV (pp. 657-663).

[3] Hahsler, Michael, and Sudheer Chelluboina. ”Visualizing association rules in hierarchical groups.” 42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms (Interface 2011). The Interface Foundation of North America. 2011.

[4] Sekhavat, Y. A., and Hoeber, O. (2013). Visualizing association rules using linked matrix, graph, and detail views.

[5] 「宇野毅明と有村博紀による公開プログラム」, http://research.nii.ac.jp/ uno/codes-j.htm, 2017/1/7 アクセス.