3O2-5

相互類似関係を用いたグラフ研磨の提案とその評価

Graph polishing technique using Mutual similarity relationship and evaluation

中原 孝信 *1 Takanobu Nakahara　岩﨑 幸子 *2 Satchiko Iwasaki　中元 政一 *2 Masakazu Nakamoto　宇野 毅明 *3 Takeaki Uno　羽室 行信 *2 Yukinobu Hamuro

*1専修大学 商学部 School of Commerce, Senshu University

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

*3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

The recent development of information technology has made bigdata analysis more familiar in research and industrial areas. We proposed graph polishing technique to clarify the unclear hidden dense structures in the graph reducing the noise. In this paper, we apply the new similarity measure "friend," which was proposed by to select useful association rules based on mutual-ranking measure between two items, to graph polishing technique and propose the visualization methodology using Nested Graph. In computational experiments, we indicate some features of the similarity measure and visualizing the graph using large scanning panel data.

1. はじめに

ICT の進歩によって，SNS や IoT など新しく膨大なデータが蓄積されており，積極的な利活用が始まっている．しかし，そのようなビッグデータにはノイズも多く含まれており，有用な情報を得るためには，ノイズを除去し，構造を明確化させることが重要となる．これまで我々はグラフ研磨と呼ばれるグラフ構造を明確化する方法を提案してきた [1].

本稿では，グラフ研磨で用いる新たな類似度を定義し，スキャンパネルデータに適用することで，これまで相関ルールで問題になっていた，1) 列挙されるルールが膨大になること，2) 出現頻度の高いアイテムを含んだ有用性の少ないルールばかりが列挙されるという問題点を解決する．

新たな類似度は相互類似関係を利用した方法であり，これは相関ルールの選択方法として提案された方法である [2]．その方法をグラフ研磨の類似度として利用することで，問題点 2) の出現頻度が大きいことで選ばれるルールを除くことが期待される．計算実験では新たな類似度指標を評価し特徴を明らかにする．

2. 手法

本稿では，グラフ研磨手法をスキャンパネルデータに適用することを前提に説明を行うが，グラフの頂点と枝が定義できれば他のデータにも適用可能である．スキャンパネルデータは，購入した商品をモニター自らがスキャンすることで収集されたデータである．トランザクション集合を D，アイテム集合を I とすると，顧客の 1 回の購買 (レシート) はトランザクション $T \subseteq D$ であり，トランザクションに出現する商品をアイテムとして，アイテム間の類似度を計算する．そして類似度がある閾値以上であればアイテム間に枝を張り，類似度グラフ $G = (V, E)$ を構成する．E はある閾値以上の類似度を持った有向枝集合で，$V \subseteq I$ はその頂点集合である．

類似度の指標としては様々なものを定義できる．例えば全トランザクションに対する共起頻度の割合である support は，式 (1) の通り定義される．ここで $\mathrm{occ}(u)$ はアイテム u が出現するトランザクション集合である．次に，confidence は，条件付き確率であり，式 (2) の通り定義される．最後に，jaccard は，和集合に対する共起頻度の割合であり，式 (3) の通り定義される．

$$\mathrm{support}(u, v) = \frac{|\mathrm{occ}(u) \cap \mathrm{occ}(v)|}{|D|} \tag{1}$$

$$\mathrm{confidence}(u, v) = \frac{|\mathrm{occ}(u) \cap \mathrm{occ}(v)|}{|\mathrm{occ}(u)|} \tag{2}$$

$$\mathrm{jaccard}(u, v) = \frac{|\mathrm{occ}(u) \cap \mathrm{occ}(v)|}{|\mathrm{occ}(u) \cup \mathrm{occ}(v)|} \tag{3}$$

これらの類似度指標とルール選択の条件である閾値を設けることで類似度グラフを構成する．どの指標を利用しても閾値を小さくするとより密に繋がった大きなグラフができるし，閾値を大きくするとより疎な小さなグラフになる．

次に，グラフ研磨アルゴリズムを Algorithm1 に示す．グラフ研磨は与えられた類似度グラフ G を枝の類似度にもとづき選択または削除することで，グラフを再構成する手法である．本稿で提案する新たな類似度を friend と呼ぶ．

Algorithm 1 グラフ研磨アルゴリズム

```
1: function Polish(G = (V, E), k)
2:     V : 頂点集合, E : 有向枝集合, k : 順位上限値
3:     V' = φ, E' = φ
4:     for all u ∈ V do
5:         for all v ∈ V do
6:             if frined(u, v) ≤ k then
7:                 E' = E' ∪ {(u, v)}
8:                 V' = V' ∪ {u, v}
9:             end if
10:        end for
11:    end for
12:    return G' = (V', E')
13: end function
```

friend は，類似度をランクに基づき評価する方法で，枝の選択方法に「相互類似関係」と「片側類似関係」の 2 種類がある．

連絡先: 中原孝信，専修大学，川崎市多摩区東三田 2-1-1，TEL:044-911-0559，nakapara@isc.senshu-u.ac.jp

それらで利用される類似度を用いたランクを式 (4) に定義する．この式により枝の評価順位が得られる．ここで $\mathrm{sim}(u,v)$ は有向枝 (u,v) の類似度である．

$$\mathrm{rank}(u,v) = |\{i : \mathrm{sim}(u,v) \leq \mathrm{sim}(u,i)\}_{i \in V}| \qquad (4)$$

次に，ランクを利用した相互類似関係による枝の選択方法を式 (5) に示す．(u,v) と (v,u) を比較し，高い方の順位が返される．そして，その値が順位上限値 k 以下であれば枝が選択され，そうでなければ削除される．この方法で選択される枝は，相互に高い類似度を持つことを意味する．

$$\mathrm{friend}(u,v) = \max(\mathrm{rank}(u,v), \mathrm{rank}(v,u)) \qquad (5)$$

次に片側類似関係を式 (6) に示す．片側類似関係では順位が k 以下なら枝が選択される．

$$\mathrm{friend}(u,v) = \mathrm{rank}(u,v) \qquad (6)$$

これまでに提案したグラフ研磨 [1] では，$\mathrm{sim}(u,v)$ と類似度の下限値 σ を利用して，$\mathrm{sim}(u,v) \geq \sigma$ なら枝 (u,v) を追加し，そうでなければ枝 (u,v) を削除することでグラフを再構築していた [*1]．つまり，アルゴリズムについては，6 行目の類似度を $\mathrm{friend}(u,v)$ または $\mathrm{sim}(u,v)$ のどちらを利用するのか，それに対応して利用される閾値が k か σ かという違いである．

最終的に Algorithm1 を繰り返し適用し，グラフ構造に変化がなくなるまで実行する．本稿では，繰り返しごとに再構成されたグラフを可視化する方法として，Nested graph[3] を利用した階層的グラフを提案する．Nested graph を利用することで，各ステップで接続のある頂点同士を 1 つのクラスタとしてまとめ，新たに生成したクラスタを次のアイテムとして利用し，階層的なグラフの可視化を行う．

Nested graph の集合を $R = \{G'_1, G'_2, \ldots, G'_n\}$ とする．n は収束するまでの繰り返し回数を表し，各繰り返しをステップと呼ぶ．G'_t はステップ t で実行した Algorithm1 の実行結果である．本稿で提案する階層的グラフでは，G'_t の連結成分を新たなアイテムとしてトランザクションデータを再構成し，連結成分をアイテムとする Nested graph を構築するところがポイントである．Algorithm2 に Nested Graph を利用した階層的グラフの構築方法を示す．

Algorithm 2 階層的グラフアルゴリズム

1: **function** HierarchyGraph($I, D, minSup, k$)
2: I : アイテム集合, D : トランザクション集合, $minSup$: サポート下限値, k : 順位上限値
3: $t = 1, D_t = D, I_t = I, R = \phi$
4: **do**
5: G_t =SimilarGraph($D_t, minSup$)
6: G'_t =Polish(G_t, k)
7: $R = R \cup \{G'_t\}$
8: $f : I_t \to I_{t+1}$ =Components(G'_t)
9: D_{t+1} =Convert(f, D_t)
10: $t = t + 1$
11: **while** $D_t \neq D_{t-1}$
12: **return** R
13: **end function**

Algorithm2 は 4 つの関数があり，SimilarGraph は，類似度グラフを生成するための関数で，トランザクション集合 D_t からアイテム間の類似度である support を計算する．そして support が $minSup$ 以上であればアイテム間に枝を張り，類似度グラフ G_t を返す．Polish は前述の通り，枝の接続と削除を行う関数で，与えられた類似度グラフから再構築されたグラフ $G'_t = (V', E')$ を返す．そのグラフの連結成分を計算するのが Components であり，G'_t から，アイテム $i \in I_t$ の属する連結成分 $i' \in I_{t+1}$ を求め，その対応関係を示すマッピング $f : I_t \to I_{t+1}$ を求める．そして，Convert ではマッピング f に従い，トランザクションデータ D_t を変換し D_{t+1} を構成する．孤立した頂点については，I_{t+1} に出現しないためトランザクションデータに出現するアイテム I_t をそのまま利用する．

このステップをトランザクションデータの更新が終わるまで続けることで，各ステップで連結成分を頂点とする Nested graph が構築できる．

3. 計算実験

計算実験のデータは (株) マクロミルより提供された QPR データ (2012/1/1〜2014/6/30) を利用し，そこから総合スーパーのイオンを選択した．モニタ数は約 2 万人，レコード数は約 380 万件で，アイテムは JICFS の細分類（984 種類）を利用した．

類似度グラフを作成するための類似度は，support を利用し最小支持度を 0.01%，と 0.005%の 2 種類，friend で利用する類似度は confidence，jaccard をそれぞれ利用し，順位の上限値となる k を 1,3,5,7,10 の 5 種類で相互類似関係による枝の選択を行った．

図 1 にその結果を示す．これは最初の 1 ステップだけを実行した結果である．

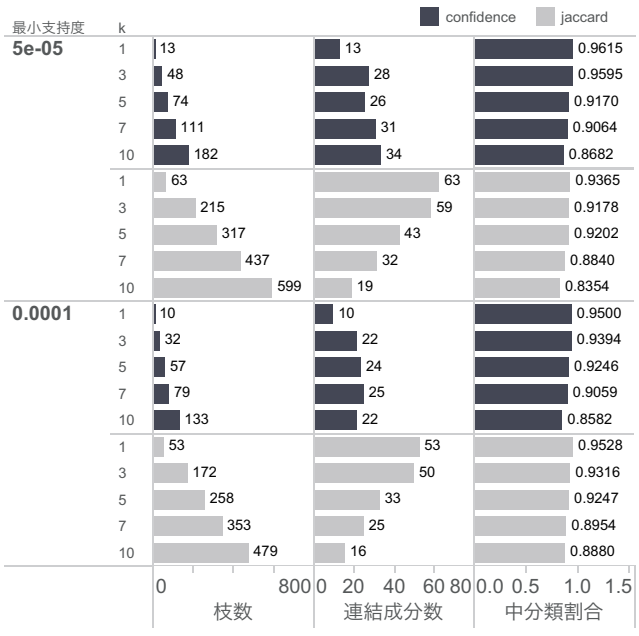

図 1: 実験の結果

濃い棒の色は confidence，薄い棒の色が jaccard による結果である．枝数は研磨後のグラフの枝数を示す．研磨前の枝数は 0.005%の最小支持度で 60,648 枝であった．jaccard の枝数は confidence に比べ多い．これは類似度の特徴で，分子の共起頻度は共通しているが分母が異なっており，confidence はお互いのランクを評価する際に，「牛乳 → きな粉」「きな粉 → 牛乳」の双方向で confidence の値が異なり，牛乳のように一方の出現頻度が大きい場合の confidence では，相互の順位が

*1 ただし $\mathrm{sim}(u,v)$ は無向枝に対して定義されていた．

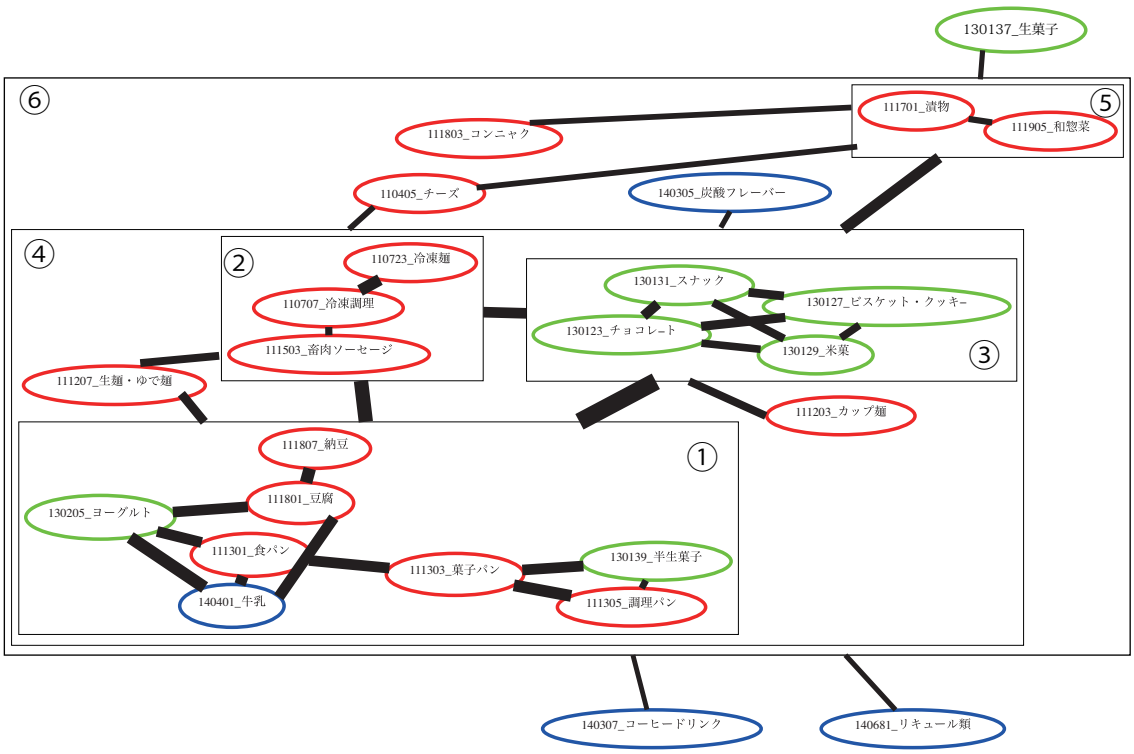

図 2: 階層的グラフ．イオンを対象に細分類を利用し，$minSup = 0.01$，類似度=jaccard，$k = 3$ で実行した結果から構築されたグラフであり，最終的に得られた 7 つの連結成分の中から食品関係の連結成分を抽出した．四角の枠はノードから構成されたクラスタを示しており，所属規則に従いマッピングすることで生成された．例えば⑤のクラスタは，マッピング規則 $f = \{(\text{漬物}, ⑤), (\text{和惣菜}, ⑤)\}$ に従う．

高くなるケースは少なくなるため接続される枝も少なくなる．jaccard は両方の頂点ペアの出現頻度が比較的等しいときに高い値を取りやすく，一方の出現頻度が大きい場合には，jaccard は小さくなる．これらの値から今回利用したデータでは，同程度の売り上げ規模を持つ商品が多い可能性がある．

連結成分数は，クラスタの数を表しており，jaccard は k を大きくすると枝のつながりが密になり，より大きなサイズのクラスタができることで，クラスタ数が少なくなっている．一方で confidence は，k が大きくなると枝の数が増えているにもかかわらず連結成分数に大きな変化はない．つまり k が大きくなったときに，これまでと異なるクラスタになるような接続が生じるわけではなく，同一クラスタ内の枝が追加されて密なグラフになっていることが考えられる．これは多様性という観点からは k が大きくなっても confidence は jaccard よりも多様性が少なくなる傾向にあると考えられる．

中分類割合は連結成分に含まれる中分類の割合をそれぞれ示したもので，複数の連結成分から最大値を選択し，それを平均した値である．この値が高ければ共通する中分類を多く含んだクラスタであることを示しており，同質性を意味している．confidence と jaccard ともに k が大きくなっても中分類割合は 80%以上であり，相互類似関係を用いることで同質性の高いクラスタが作成できている．これら 3 つの値からある程度広いつながりを持った網羅性の高いグラフを抽出したければ，confidence よりも jaccard を利用することが好ましい．

表 1 は，相互類似関係と片側類似関係を比較した結果である．

表 1: 頻度差のあるルールの割合

方向	support	類似度指標	頻度差割合
相互	0.01%	jaccard	3.0%
相互	0.005%	jaccard	2.5%
相互	0.01%	confidence	9.2%
相互	0.005%	confidence	8.7%
片側	0.01%	jaccard	21.2%
片側	0.005%	jaccard	18.5%
片側	0.01%	confidence	87.1%
片側	0.005%	confidence	85.9%

「方向」の相互は式 (5) による枝の選択，片側は式 (6) で選択した枝による結果である．出現頻度の高い上位 10 アイテムを高出現アイテムとし，u か v のいずれか一方のみが高出現アイテムの割合を示したものが「頻度差割合」である．

頻度差を伴った枝の割合が高い場合は，(u, v) のどちらか一方の頂点が高出現アイテムによって選択されていることを意味しており，商品間の共起頻度が一方の商品の大きさに依存して高くなっていることを示している．相互類似関係では一方が高出現アイテムの枝は比較的少なく，また，類似度の特徴としては confidence よりも jaccard を用いた場合は更に少なくなっている．このことから 1. はじめにで示した相関ルールの問題点 2) に対する解決策の 1 つになることが考えられる．

図 2 は，階層的グラフから抽出した 1 つの連結成分を示している．イオンを対象に細分類を利用し，$minSup = 0.01$，類似度=jaccard，$k = 3$ で実行した結果から構築されたグラフで，最終的に 7 個の連結成分ができており，図は食品関係の連結成分を抽出したものである．

①は食パン，菓子パン，牛乳，豆腐，ヨーグルトなど乳製品とパンとのつながりを示したクラスタで，売り上げの大きい商品同士が双方向に繋がっており朝食を連想させるようなクラスタである．②は畜肉ソーセージ，冷凍食品で弁当のおかずを連想させるようなクラスタである．③はお菓子のクラスタで，これら 3 つのクラスタ同士が互いに接続し，生麺・ゆで麺，カップ麺など麺類を表す単一のアイテムとともにクラスタ④を構成している．

これらの関係からクラスタ④は，売り上げの大きな加工食品クラスタであり，朝食，弁当，お菓子などの関係性から主婦の典型的な購買行動の 1 つであると解釈できる．そして，このクラスタと和惣菜と漬物からなるサイドメニューのクラスタ⑤が相互に接続し，チーズ，炭酸フレーバー，コンニャクなどとクラスタ⑥を構成している．そして最後にそれらのクラスタに対して，コーヒー・ドリンク，リキュール類，生菓子が接続されている．

階層的グラフを利用することでアイテム間の関係性からクラスタの解釈ができ，それと同時にクラスタ同士の関係性も理解することができる．

4. おわりに

本稿では，相関ルールの選択方法として提案された friend をグラフ研磨の新たな類似度として利用する方法とその特徴を示した．計算実験からは confidence と比較し jaccard を用いることで，アイテムの関係性を俯瞰できる広いつながりと，多様性のあるアイテム間のつながりを捉えられることを示した．また，相互類似関係を利用することで，相関ルールで問題になりやすい，片方のアイテムが大きいことによって抽出されるルールを改善できることを示した．そして Nested Graph を利用した階層的グラフの可視化方法では，アイテム間の関係性からクラスタの解釈ができること，同時にクラスタ同士の関係性も理解できることを示した．

謝辞

(株) マクロミルよりスキャンパネルデータ QPR を提供いただいた．本研究の一部は，科学技術振興機構 CREST の研究助成，また JSPS 科研費 JP15K17146 の助成を受けたものである．

参考文献

[1] 宇野毅明, 中原孝信, 前川浩基, 羽室行信, データ研磨によるクリーク列挙クラスタリング, 情報処理学会 AL 研究会報告書, 2014-AL-146(2), pp. 1-8, 2014.

[2] 岩崎幸子, 中元政一, 中原孝信, 宇野毅明, 羽室行信, グラフ構造による相関ルールの視覚化ツール:KIZUNA, 日本 OR 学会 2017 年春季全国大会,2017.

[3] Alexandra Poulovassilis and Mark Levene. A nested-graph model for the representation and manipulation of complex objects. ACM Trans. Inf. Syst. 12, 1 (January 1994), 35-68. DOI=10.1145/174608.174610 http://doi.acm.org/10.1145/174608.174610