大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

数理システム　ユーザーコンファレンス2018

クラスタリングの安定化と相互類似関係を考慮したアソシエーション分析
～データマイニング手法の困った問題への対処案～

2018年11月22日

情報・システム研究機構
国立情報学研究所
情報学プリンシプル研究系

岩崎　幸子

本日の内容

1.NIIと研究プロジェクトの紹介

2.データマイニング手法の困った問題

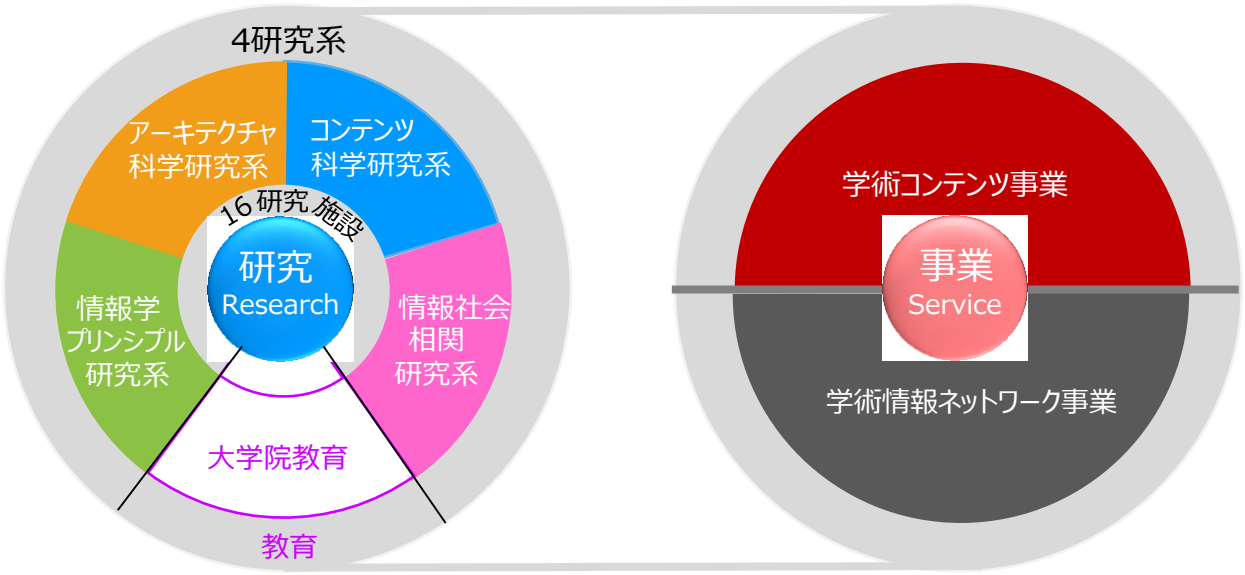
3.クラスタリングの安定化（コンセンサス・クラスタリング）

4.相互類似関係を考慮したアソシエーション分析

1.NIIと研究プロジェクトの紹介

国立情報学研究所（NII）のご紹介

弊所は国内唯一の情報学の学術総合研究所です。情報学における基礎論から、人工知能、情報セキュリティなどの幅広い研究分野において、長期的な視点に立つ基礎研究から社会課題の解決を目指した実践的研究を推進しています。
また、大学共同利用機関として、学術コミュニティ全体の研究・教育活動に不可欠な学術情報基盤の構築、学術コンテンツとサービスプラットフォームの提供といった事業を展開しています。



情報から知を紡ぎだす
研究と事業を両輪として、情報学による未来価値を創成します

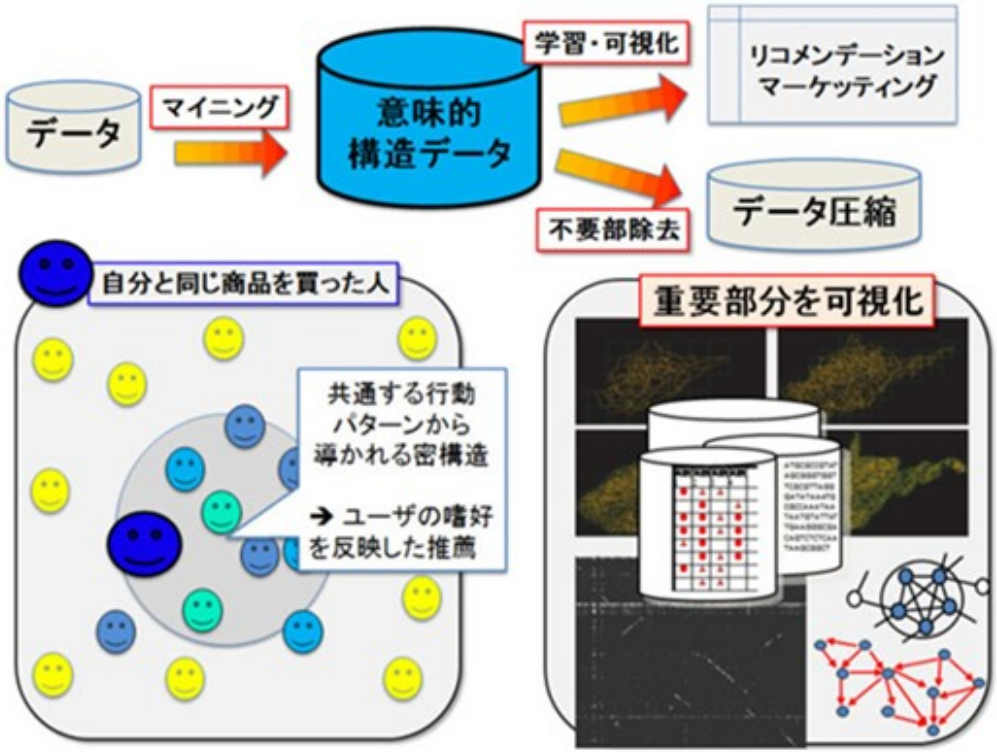
参画する研究プロジェクト（JST CREST）

ビッグデータ統合利活用に関する、国が定める戦略目標の達成に向けて、
「データ粒子化」による高速高精度な計算技術の研究開発に取り組んでいます。
アルゴリズム開発から産業界での応用研究などを行っています。

CREST

研究代表者
宇野毅明
（国立情報学研究所教授）

データ粒子化による高速高精度な次世代マイニング技術の創出



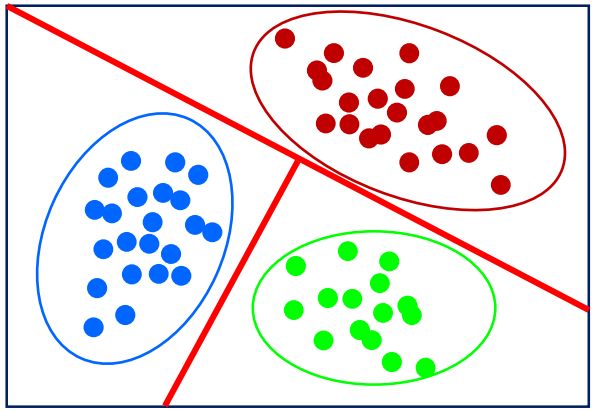
出所:ビッグデータから隠れた知識を探り出す超高速解析アルゴリズム「NII SEEDs」
http://www.nii.ac.jp/userdata/seeds/2014uno.html

2.データマイニング手法の困った問題

クラスタリングとアソシエーション分析

現在、クラスタリングとアソシエーション分析は非常に多くの方々に使われているデータマイニングの主要な手法です。企業のマーケティング活動においても、戦略立案から戦術策の立案など、様々な意思決定場面でこれらの手法を用いたデータ分析が利用されています。

クラスタリング（Clustering）



どういう手法？	利用例
データを似ている特徴をもつグループに分ける手法。購買特徴の似た顧客やテキストデータの分類などによく使われている。	顧客理解と対応 ・店舗/売場開発 ・商品開発 ・レコメンデーション広告

アソシエーション分析（Association Rule Mining）

X	→	Y	support
牛乳	→	パン	20％
パン	→	バター	5 ％
豆腐	→	納豆	
レタス	→	きゅうり	
トマト	→	レタス	
リンゴ	→	バナナ	

どういう手法？	利用例
「牛乳」を買ったら→「パン」を買うというように、ある事象が起きると別の事象が起きるという事象間の共起性のルール（相関ルール）を抽出する手法。	販売促進 ・フロアレイアウト/棚割 ・クロスセリング（併売） ・レコメンデーション広告

クラスタリングをやっていて困ること

手法によって異なりますが、メジャーな手法に多くあるのが「初期値依存」の問題。
出力結果が非常に不安定で、初期値変われば結果が様変わりすることがあります。
「精度評価値が低いモデルを選ぶ」という方法は本当に現場での解決になっているのでしょうか？

Algorithms

K-Means
K++
PLSA

主な問題

・結果が初期値に依存する。
・初期値変われば結果が
様変わりする。

一般的な対応策

複数の初期値で試行して、
精度評価値の低いモデルを選ぶ。

SSE　AIC　BIC

クラスタ	精度評価１位		精度評価２位		精度評価３位	
	解釈	比率	解釈	比率	解釈	比率
CL1	惣菜多い簡便クラスタ	36%	魚が多い和食系クラスタ	32%	野菜と果物が多いクラスタ	22%
CL2	肉多い洋食系クラスタ	22%	肉多い洋食系クラスタ	29%	魚が多い和食系クラスタ	19%
CL3	魚が多い和食系クラスタ	15%	高級食材系クラスタ	18%	肉多い洋食系クラスタ	16%
CL4	冷食とインスタントクラスタ	11%	惣菜多い簡便クラスタ	12%	お菓子多いおやつクラスタ	13%
CL5	？？？解釈不能クラスタ	6%	健康志向クラスタ	7%	惣菜多い簡便クラスタ	9%

※内部計算の学習回数については十分に考慮して実験しています。

アソシエーション分析をやっていて困ること

頻出パターンの計算は、組み合わせの数が多くなることから計算量も多く膨大なルールが列挙されがちです。対象アイテムを少なくしたり、評価指標に閾値を設定して組み合わせを削減しますが、結果には無意味なルールが多く含まれルール選択に骨が折れます。

Algorithms

Apriori
FPGrowth
LCM

主な問題

・計算に時間がかかる。
・膨大なルールが列挙される。
・無意味なルールが多く、
　選択に骨が折れる。
・アイテム全体の関係を
　俯瞰できない。

一般的な対応策

・対象アイテムを少なくする
・評価指標に閾値を設定し、
　ルールを足切りする。
・評価指標を複合的に評価して、
　ルール選択を行う。

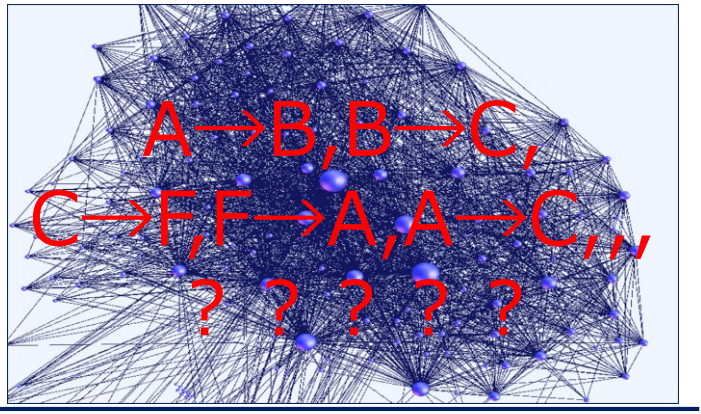
支持度　確信度　リフト

result

	前件部	小分類名称	後件部	小分類名称:0	同時購買数	Support	ConfA	ConfB	Lift
1	34	野菜加工品	327	豆腐	491753.00	2.41	25.72	19.10	2.03
2	327	豆腐	34	野菜加工品	491753.00	2.41	19.10	25.72	2.03
3	344	牛乳	327	豆腐	441186.00	2.17	21.61	17.17	1.71
4	327	豆腐	344	牛乳	441186.00	2.17	17.17	21.61	1.71
5	330	納豆	327	豆腐	428[illegible]	2.09	[illegible]0	16.54	2.81
6	327	豆腐	330	納豆	428861.00	2.09	16.54	35.50	2.81
7	47	その他飲料	327	豆腐	389160.00	1.91	37.58	15.11	2.97
8	327	豆腐	47	その他飲料	389160.00	1.91	15.11	37.58	2.97
9	384	菓子パン	327	豆腐	380674.00	1.87	14.75	[illegible]	[illegible]
10	327	豆腐	384	菓子パン	380674.00	1.87	14.75	[illegible]	[illegible]
11	902	その他	327	豆腐	367506.00	1.80	[illegible]	[illegible]	[illegible]
12	327	豆腐	902	その他	367506.00	1.80	[illegible]	[illegible]	[illegible]
13	384	菓子パン	344	牛乳	361813.00	1.78	[illegible]	[illegible]	[illegible]
14	344	牛乳	384	菓子パン	361813.00	1.78	[illegible]	[illegible]	[illegible]
15	331	漬物	327	豆腐	357081.00	1.75	[illegible]	[illegible]	[illegible]
16	327	豆腐	331	漬物	357081.00	1.75	[illegible]	[illegible]	[illegible]
17	327	豆腐	15	バナナ	311971.00	1.53	[illegible]	[illegible]	[illegible]
18	15	バナナ	327	豆腐	311971.00	1.53	[illegible]	[illegible]	[illegible]
19	38	その他菓物	327	豆腐	309140.00	1.52	[illegible]	[illegible]	[illegible]
20	327	豆腐	38	その他菓物	309140.00	1.52	[illegible]	[illegible]	[illegible]
21	328	生めん	327	豆腐	304896.00	1.50	[illegible]	[illegible]	[illegible]
22	327	豆腐	328	生めん	304896.00	1.50	[illegible]	[illegible]	[illegible]
23	349	ヨーグルト	327	豆腐	301949.00	1.48	[illegible]	[illegible]	[illegible]
24	327	豆腐	349	ヨーグルト	301949.00	[illegible]	[illegible]	[illegible]	[illegible]
25	349	ヨーグルト	344	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
26	344	牛乳	349	ヨーグルト	301949.00	[illegible]	[illegible]	[illegible]	[illegible]
27	382	食パン	344	牛乳	285499.00	[illegible]	[illegible]	[illegible]	[illegible]
28	344	牛乳	382	食パン	285499.00	[illegible]	[illegible]	[illegible]	[illegible]
29	34	野菜加工品	328	生めん	277408.00	1.36	[illegible]	[illegible]	[illegible]
30	328	生めん	34	野菜加工品	277408.00	1.36	[illegible]	[illegible]	[illegible]
31	327	豆腐	1	きゅうり	276956.00	1.36	[illegible]	[illegible]	[illegible]
32	1	きゅうり	327	豆腐	276956.00	1.36	[illegible]	[illegible]	[illegible]
33	384	菓子パン	34	野菜加工品	274891.00	1.35	[illegible]	[illegible]	[illegible]

数十万行

牛乳　→もやし？
バナナ →もやし？
もやし　→納豆？
出現頻度の高い
アイテムがたくさん。



困った問題への対処案

クラスタリングの初期値依存問題

クラスタリングの安定化（コンセンサス・クラスタリング）

・複数の初期値で得たクラスタリング結果から、
頑健なクラスタを捉えてより安定した解を得る。

膨大なアソシエーション・ルール列挙の問題

相互類似関係を考慮したアソシエーション分析

・アソシエーション分析の結果を使って、
アイテム間の親しさを考えてルール抽出を行う。

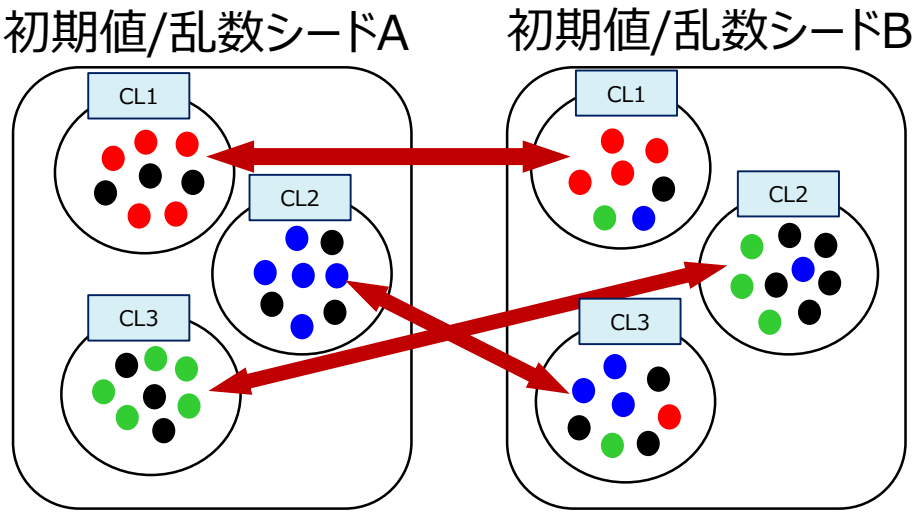
3.クラスタリングの安定化
（コンセンサス・クラスタリング）

一般的なクラスタリング手法はどのぐらい不安定か

3種類のアルゴリズムに対し、初期値を変えてそれぞれ100回のクラスタリングを実施。
初期値が異なるモデル間の、同じ人が含まれる最も類似度の高いクラスタ同士を比較しました。

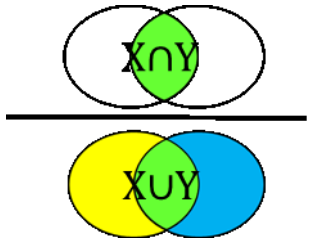
計算実験

・対象データ:QPR消費者購買履歴データ
・対象人数 :14,648人
・クラスタ数 :30
・初期値試行:100回



計算実験結果

	K-Means	K++	PLSA
類似度 jaccard	18.7%	22.8%	53.9%



出所：参考文献[1] 宇野毅明,岩﨑幸子,中原孝信,中元政一,羽室行信,
“乱数シード依存のクラスタリング手法の安定化に対するアプローチ”,
人工知能学会人工知能基本問題研究会 105,pp.58-62 (2018).

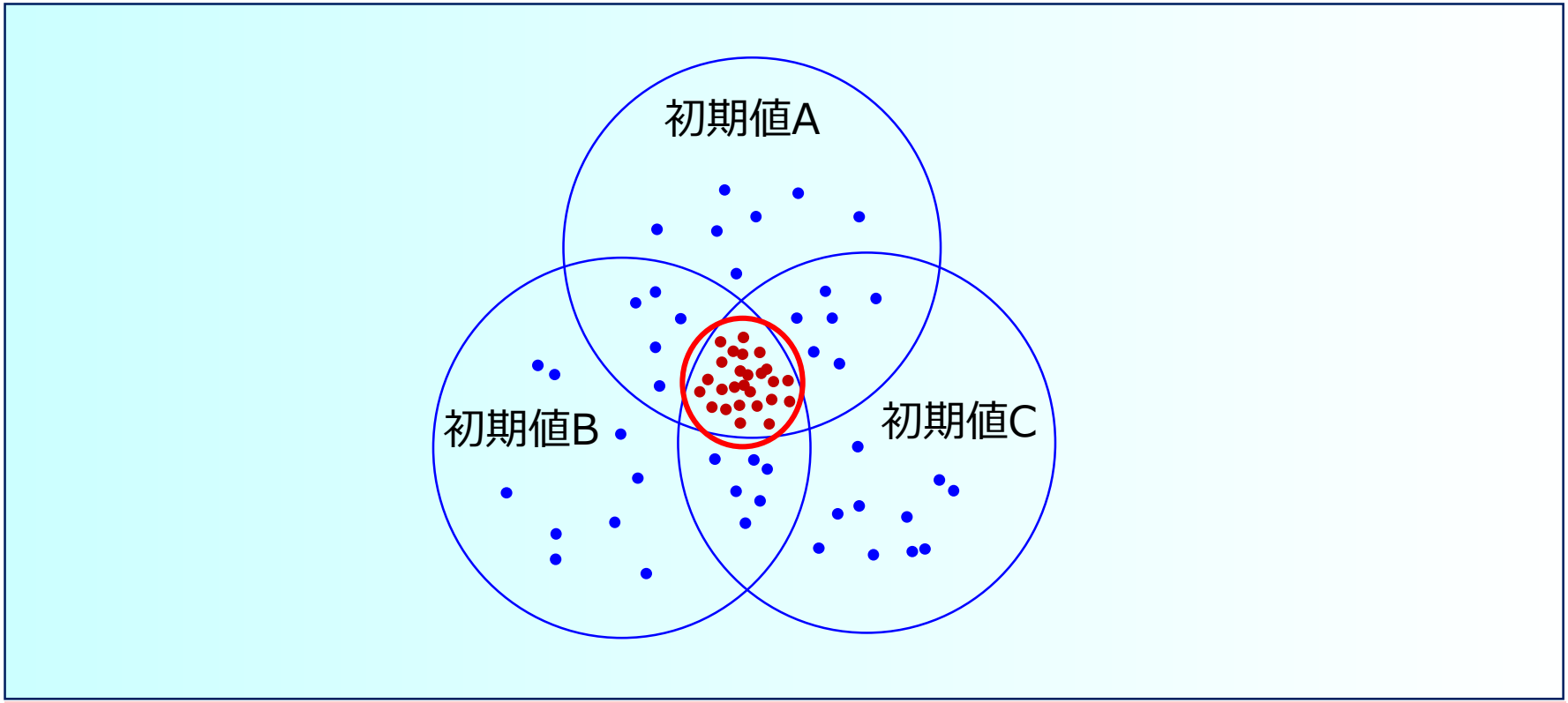
精度評価の良いモデルを比較しても、、、

試行回数1000回で精度評価上位２つのモデルを比較

	1位	２位	類似度
K-Means	seed=3	seed=654	30.4%
K++	seed=334	seed=127	38.0%
PLSA	seed=287	seed=975	68.7%

不安定なものの中にも安定しているものはある

初期値を変えて何回計算しても、必ず同じクラスタに所属する頑健な集合「芯クラスタ」があります。結果を不安定にしているのは「芯クラスタ」以外の部分になります。



まずは頑健な「芯クラスタ」を捉え、そして不安定なものをどう扱うかを、分析目的に応じて考えると良いのではないか？

クラスタリングの安定化（コンセンサス・クラスタリング）とは？

複数回のクラスタリングの結果を使って新しいクラスタを生成するアプローチは、

・コンセンサス・クラスタリング
・アンサンブル・クラスタリング

※基本的に精度評価値を向上させることが目的の研究

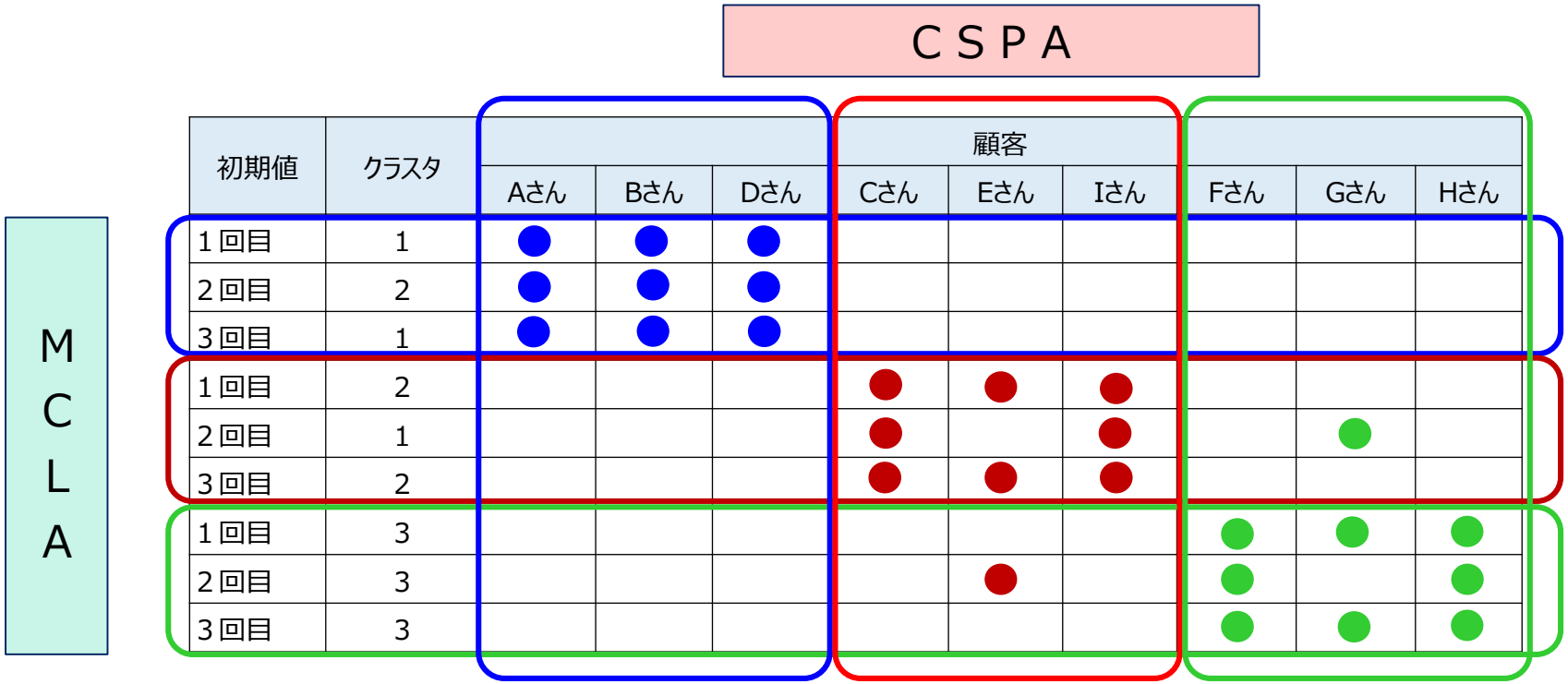
弊所は、

信頼性の高い解を得ることを目的に、コンセンサスのとれた安定した解がほしい。
本来、アルゴリズムが出したいであろう解にできるだけ近づきたい。

クラスタリング安定化のアプローチ

複数の初期値にて実施したクラスタリング結果から、「芯クラスタ」を捉えます。
それには２つのアプローチがあります。

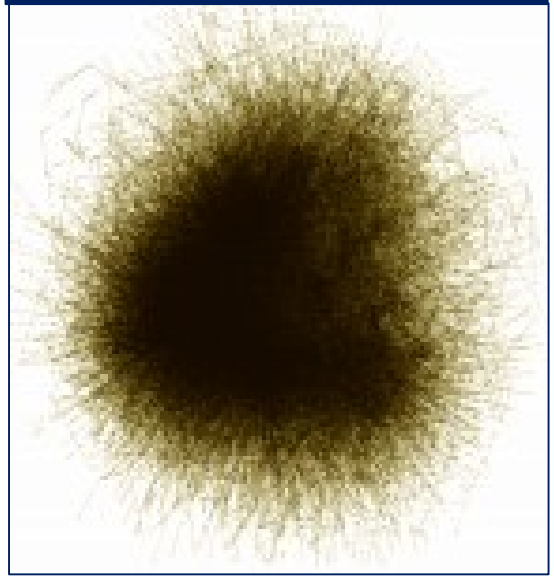
1. ＭＣＬＡ （Meta-CLustering Algorithm）
2. ＣＳＰＡ （Cluster-based Similarity Partitioning Algorithm）



データ粒子化技術の適用（データ研磨による再クラスタリング）

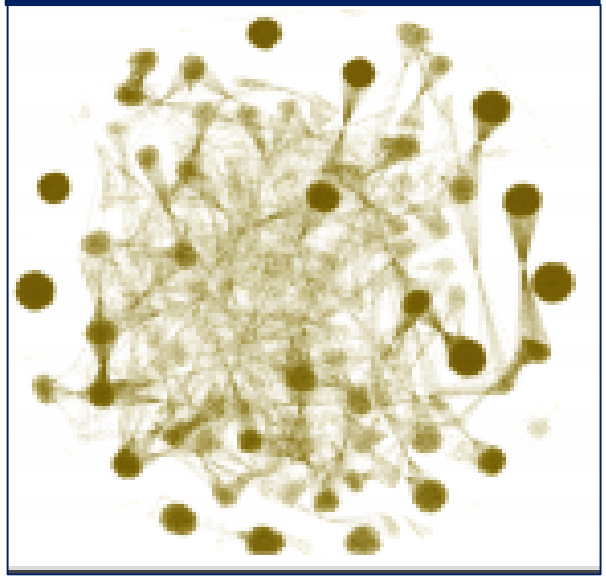
データ研磨（Data Polishing）というマイクロ・クラスタリングの手法があります。
データの大まかな構造を変えることなく、強い関係は強く、弱い関係は弱くすることで、
類似した性質を持つグループを抽出します。頑健で細やかなグループを抽出することが可能です。

元の複雑データ



Polish!

研磨後のデータ



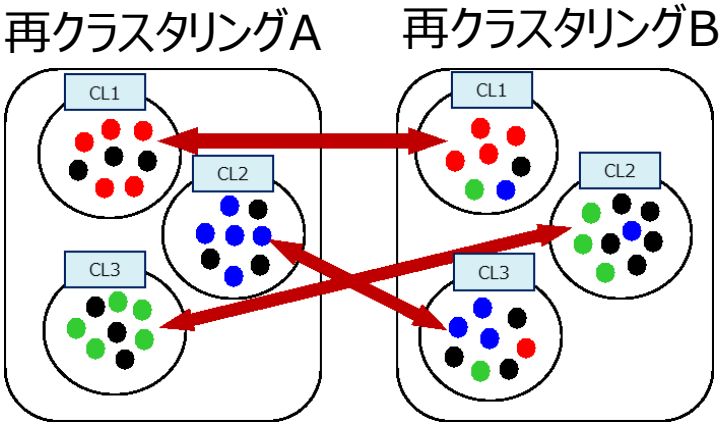
NYSOLはオープンソース（無料）です。
大規模データの解析に関する様々な研究成果を、
広く産業界に還元する目的で構築しています。

https://www.nysol.jp/

安定化実験の結果

計算実験

・対象データ:QPR消費者購買履歴データ
・対象人数 :14,648人
・クラスタ数 :30
・初期値試行:100回



	K-Means	K++	PLSA
元のクラスタリングの類似度	18.7%	22.8%	53.9%

		K-Means	K++	PLSA
MCLA	5回分	44.7%	45.0%	63.8%
	10回分	45.9%	50.7%	69.6%
	20回分	49.6%	48.0%	73.4%
	50回分	57.8%	57.8%	74.4%
CSPA	5回分	32.6%	32.4%	72.2%
	10回分	42.2%	42.0%	60.0%
	20回分	50.4%	46.8%	74.3%
	50回分	55.8%	53.9%	80.4%

出所：参考文献[2] 宇野毅明,岩﨑幸子,中原孝信,中元政一,羽室行信,
“データ研磨によるコンセンサスクラスタリングの精緻化”,人工知能学会人工知能基本問題研究会 106,pp.43-50 (2018).

本日ご紹介する企業事例

① PLSAを用いた食品スーパーの顧客コンセンサス・クラスタリング

② 商品と店舗の改善余地（伸びしろ）の推定

発表のみ

4.相互類似関係を考慮したアソシエーション分析

CRESTプロジェクトの取組み

アソシエーション分析の問題

1.ルール計算に時間がかかる

2.膨大なルールが列挙され、
無意味なルールが多くフィルタリングに
骨が折れる

3.データリストを読むだけでは、
複数のルール間の関係を把握し、
全体を俯瞰することが困難

世界最高速の列挙アルゴリズム
SSPC(Similar Set Pair Comparison)

本提案手法によるルール選択

ネットワークグラフによる視覚化

KIZUNA

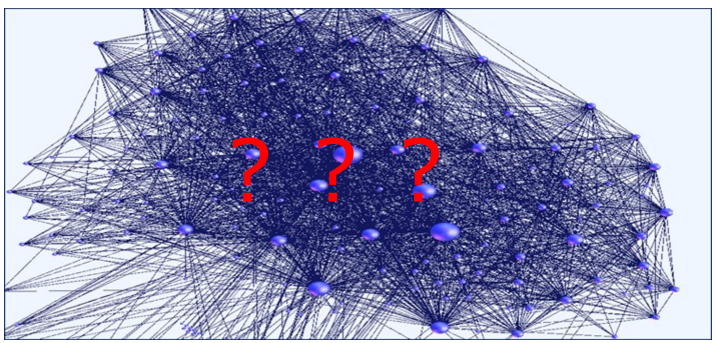
アプリケーションツールに実装

POLISH FRIEND PAL

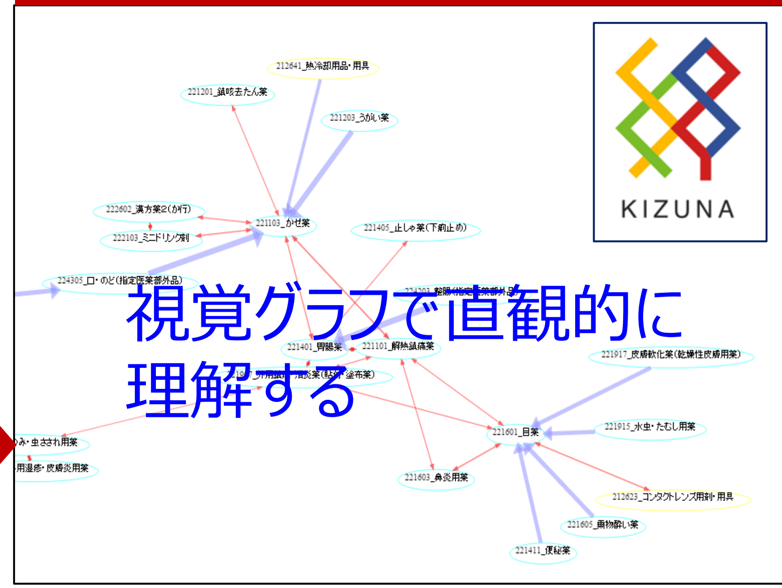
効率的な分析アプローチを提案

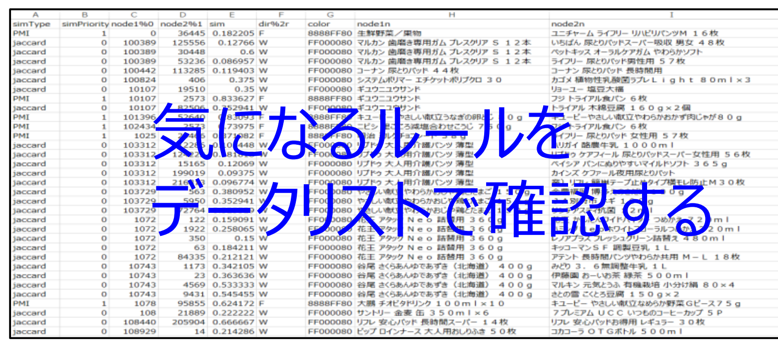
NII

一般的なアプローチ



提案アプローチ





目的は分析をすることではなく、
結果から打ち手を考えアクションを起こすこと

提案手法

- トランザクションデータから列挙される膨大なアソシエーション・ルールから、少数の有用である可能性の高いルールを選択する手法です。
  必要に応じてグラフ描画を行います。

STEP1 アソシエーション分析

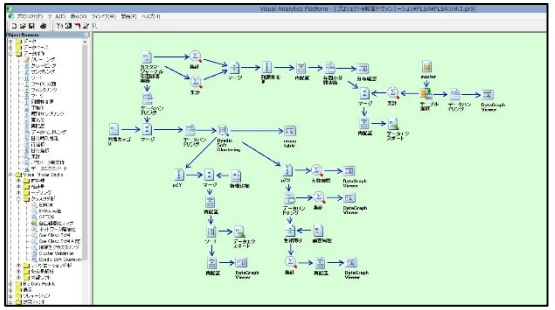
↓

STEP2 相互類似関係によるルール選択

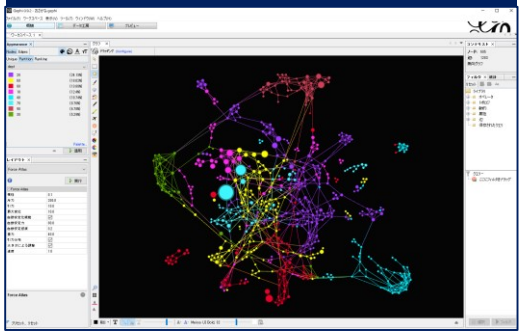
↓

STEP3 グラフ構造による視覚化

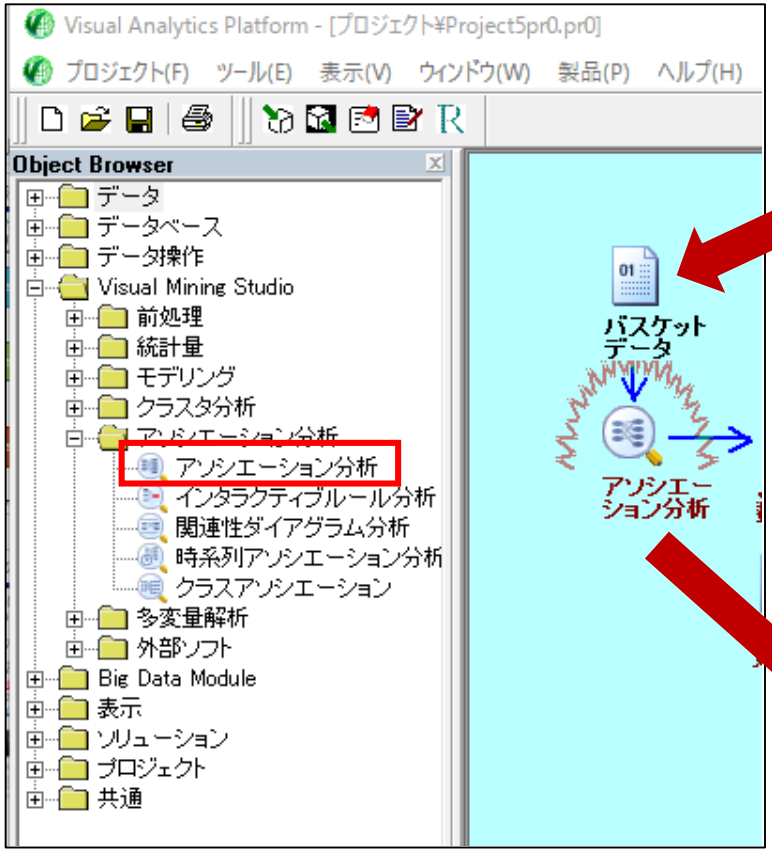
Visual Mining Studio®
汎用データマイニングシステム



グラフ描画ツール



トランザクションデータに対しルール評価指標の下限値を与え、その条件を満たすルール集合を出力します。



バスケットデータ --- データ＆グラフビュー

table　レシートや人単位に商品などを集計したデータ

	月	レシート番号	カテゴリコード	点数
1	4	1235	87	1.0000000000
2	4	1235	233	1.0000000000
3	4	1235	524	1.0000000000
4	4	1235	565	1.0000000000
5	4	1235	614	1.0000000000
6	4	1236	408	1.0000000000
7	4	1236	512	1.0000000000
8	4	1236	595	1.0000000000
9	4	1236	614	1.0000000000
10	4	1236	851	1.0000000000
11	4	1237	310	1.0000000000

テーブル：19,864,618 行 5 列

アソシエーション分析 --- データ＆グラフビュー

result　アソシエーション・ルール集合

	前提	結論	信頼度	サポート	Lift
1	818	506	100.0000000000	0.0000658828	21.6039027584
2	780	1	75.0000000000	0.0000988242	5.0761623733
3	780	44	75.0000000000	0.0000988242	4.0877867529
4	62	44	66.1075367647	0.0947723980	3.6031135074
5	61	44	57.1784316715	1.1118709593	3.1164431405
6	55	44	52.9924884904	0.1440856687	2.8882932328
7	134	614	51.6772438803	0.0187765961	4.1329569250
8	210	455	51.0050806273	0.0760616847	9.1295780157
9	686	44	50.6911240546	0.1920812836	2.7628600720
10	509	44	50.1448645516	1.5963730205	2.7330868406
11	830	829	50.0315457413	0.0261225275	19.6835789087

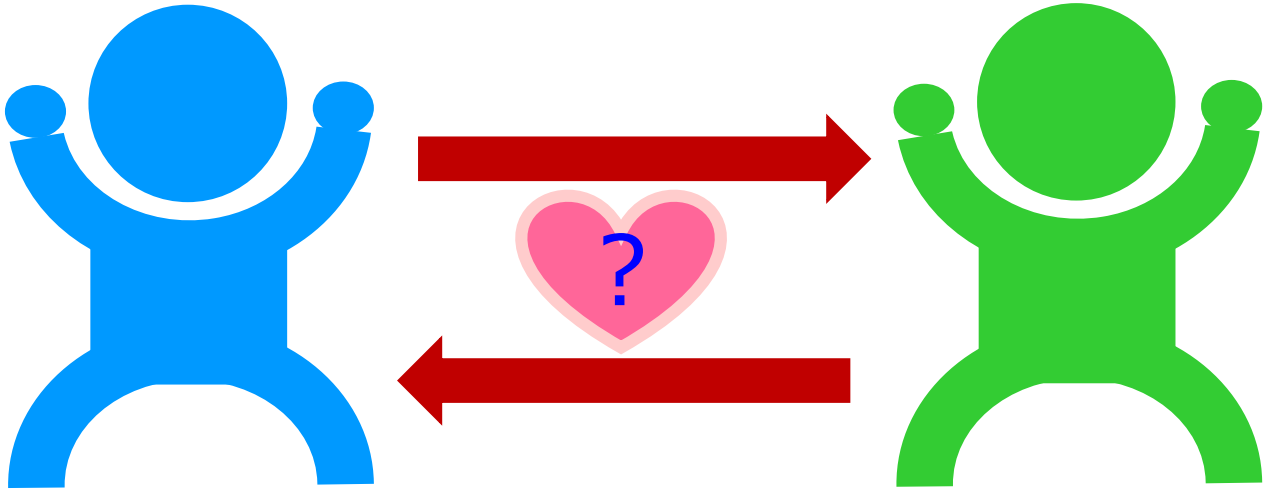
テーブル：395,564 行 10 列

$X \Rightarrow Y$

STEP2：相互類似関係によるルール選択

基本発想

お互いにとっての「親しさ」を考慮することで、何らかの意味を持つ有用なルールがより多く抽出でき、情報量を削減できないだろうか？



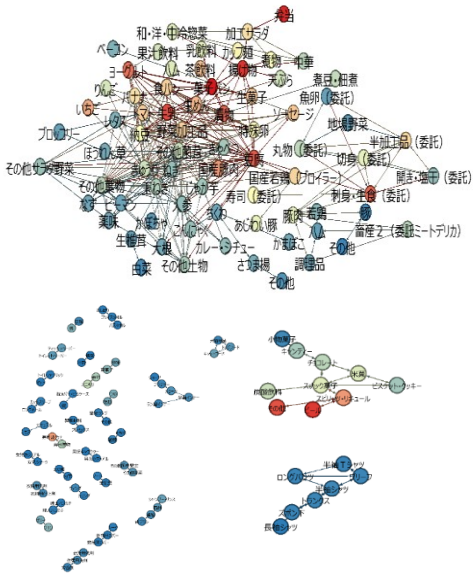
お互いにとっての「親しさ」を何ではかるか？

ランク情報を用いることで視覚化もしやすくなる

多くの視覚化に共通した問題点

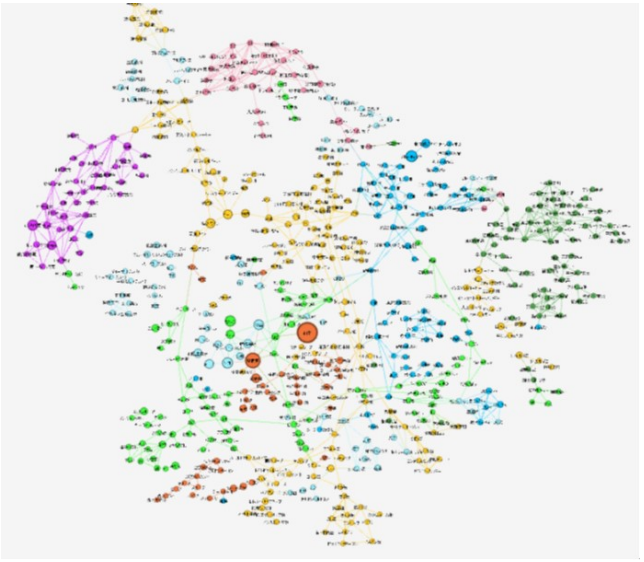
個別ルールの把握（ローカル性）とルールの全体的な関係性の俯瞰（グローバル性）を同時に実現することが困難

閾値での枝狩り



ネットワークグラフの枝が複雑になる。
解消のために閾値を上げるほどに、
アイテムが減って残らなくなる。

ランク情報による枝狩り



アイテムの次数（つながりの数）に
制限を入れることができる。
複雑化を回避し、全てのアイテムを
表現することが可能になる。

網羅性〇
連結性〇
分散性〇

網羅性：アイテムカバー率
（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：連結成分数
（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：最大次数
（特定のアイテムに接続が極度に
集中しない）

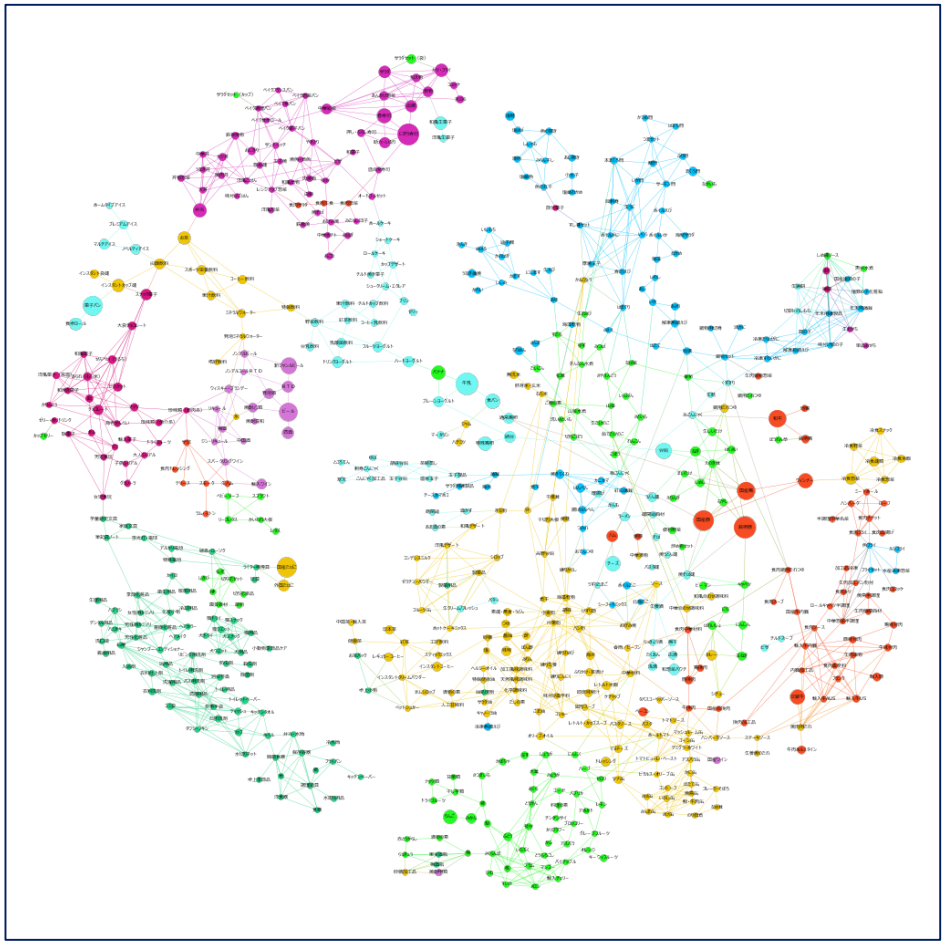
企業事例：食品スーパーのバスケット分析

クロスMD（関連販売）や売場配置への知見を得るために、
興味深いルールを発見することを目的としました。

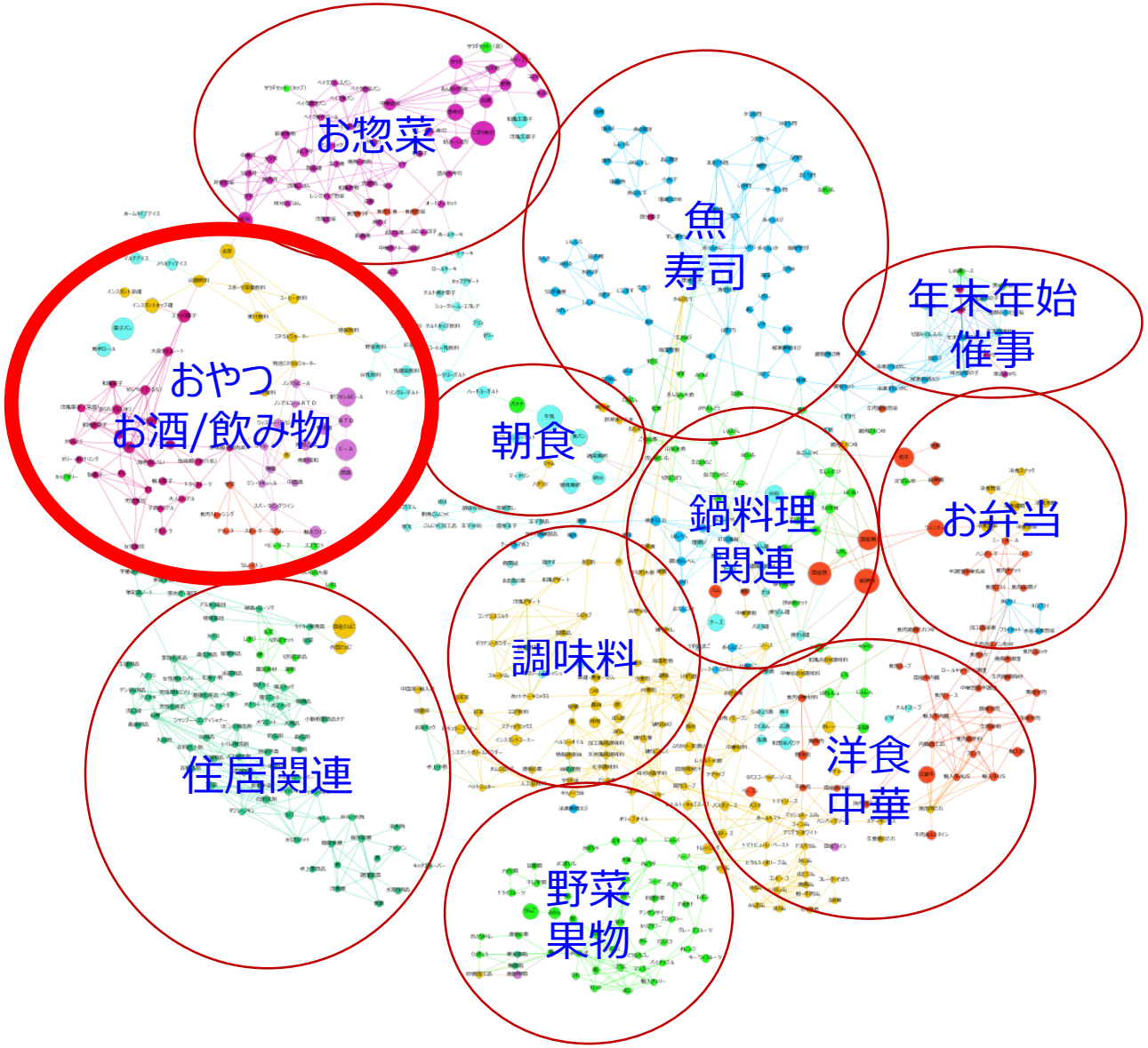
小売業A社
SM/SSM業態

❤両思い複合グラフ
ルール条件：

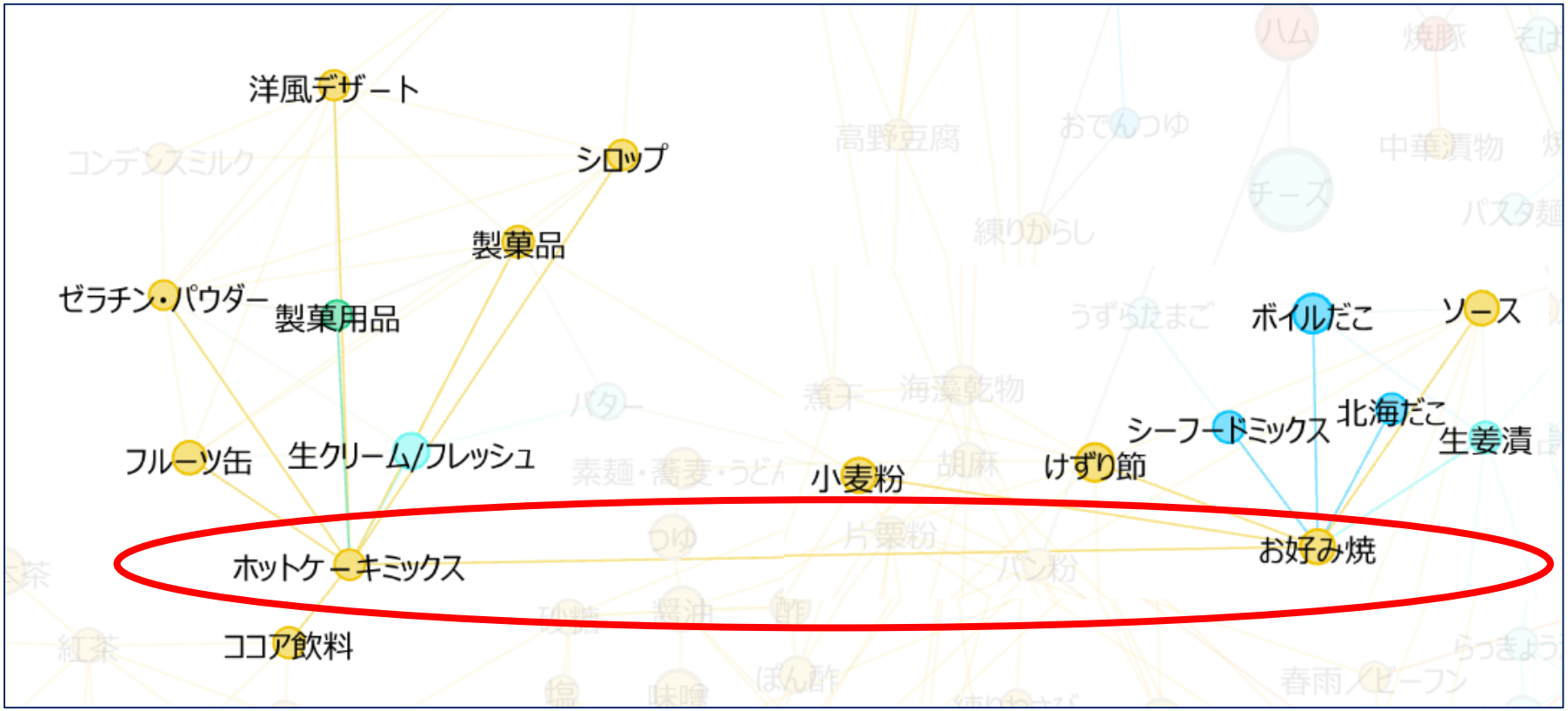
ノードの大きさ：売上金額
ノードの色　　：部門分類



全体のネットワーク



ホットケーキとお好み焼き？



まとめ

・クラスタリングとアソシエーション分析の問題として、
クラスタリングの初期値依存問題と膨大なルール列挙の問題をあげました。

・クラスタリングの初期値依存問題には、コンセンサス・クラスタリングを用いることで、
シングル・クラスタリングよりも安定し、意味解釈可能な信頼性の高い結果を
得られる可能性があります。

・アソシエーション分析の膨大なルール列挙には、相互類似関係を考慮し、
ランク情報によるルール選択を行うことで意味解釈ができる結果が得られます。

本提案手法が皆さまのご参考になると幸甚です。

ご清聴ありがとうございました。

参考文献

[1] 宇野毅明,岩﨑幸子,中原孝信,中元政一,羽室行信,
“乱数シード依存のクラスタリング手法の安定化に対するアプローチ”,
人工知能学会人工知能基本問題研究会 105,pp.58-62 (2018).

[2] 宇野毅明,岩﨑幸子,中原孝信,中元政一,羽室行信,
“データ研磨によるコンセンサスクラスタリングの精緻化”,
人工知能学会人工知能基本問題研究会 106,pp.43-50 (2018).

[3] Takeaki Uno, Hiroki Maegawa, Takanobu Nakahara, Yukinobu Hamuro,
Ryo Yoshinaka, and Makoto Tatsuta,
“Micro-Clustering by Data Polishing, ” IEEE Big Data 2017 (2017).

[4] 岩﨑幸子,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”,
人工知能学会全国大会論文集, 2L42 (2017).

[5] 中原孝信, 岩﨑幸子, 中元政一,宇野毅明, 羽室行信,
“相互類似関係を考慮したグラフ研磨の提案とその評価”,
人工知能学会全国大会論文集, 3O25 (2017).

発表内容に関するお問い合わせは、iwasaki@nii.ac.jp にご連絡ください。