10 回ずつ BIC の値を調べた。その結果、潜在クラス数は平均的に 20 が最適であると判断されたため、その値を用いる。潜在クラス数 20 と設定した PLSI を各 30 回実行し、最も尤度が高かった分類結果を採用した。その分類の結果の商品の傾向から各 20 カテゴリに対し名前を付与した。その名前、代表的な商品例、分類顧客数、分類商品数を表 2.2.1-3 に示す。この名前の付与は便宜上与えたもので本質的に必要なものではない。また、図 2.2.1-8 にライフスタイルカテゴリと新しい商品カテゴリの関係を示す。図中の実線が、各商品カテゴリ内での最も得点が高いライフスタイルカテゴリであり、破線が各ライフスタイルカテゴリに対しての新しい商品カテゴリでの得点が高い方から 3 カテゴリを結んだ図である。また、太線は両方の結びつきがある線を表す。

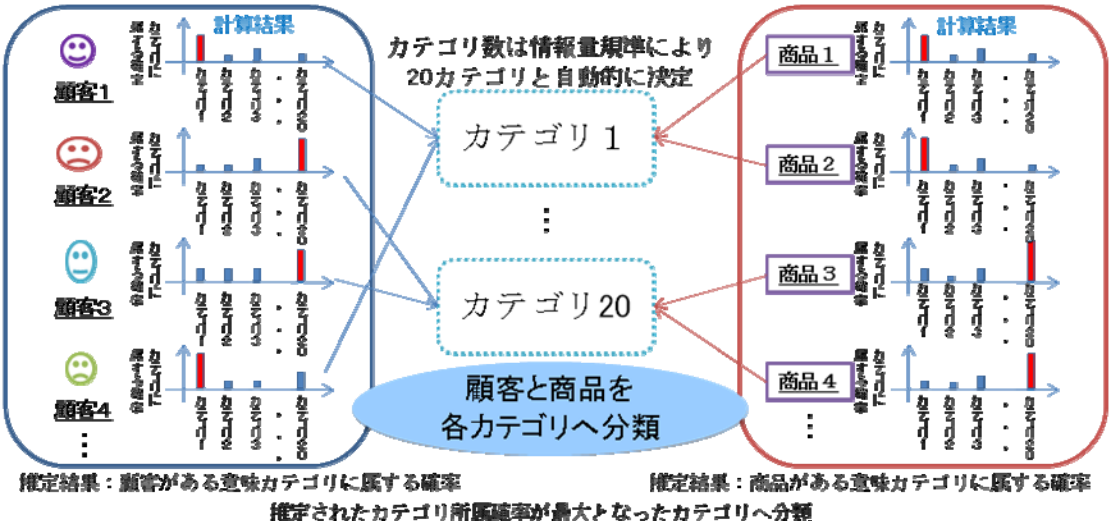

図 2.2.1-7　確率的潜在意味解析の概念図

表 2.2.1-3　PLSI による分類結果

クラスタ番号	名前	代表的な商品	分類顧客数	分類商品数
クラスタ番号１	果物自炊的	いちご、りんご、みかん、だいこん、魚	469	81
クラスタ番号２	お手軽夕食的	コロッケ、天ぷら、巻きずし、唐揚げ	185	94
クラスタ番号３	酒飲み健康的	ビール、発泡酒、乳製品	39	21
クラスタ番号４	パン食的	食パン、イチゴジャム、チョコクリーム	95	10
クラスタ番号５	野菜自炊的	もやし、えのき、きゅうり、たまねぎ	495	76
クラスタ番号６	おやつ的	菓子パン、団子、プリン、ヨーグルト	109	73
クラスタ番号７	洋風朝食的	牛乳、ハム、コーヒー	39	6
クラスタ番号８	牛乳・清涼飲料的	牛乳、ペットボトルの水、お茶、コーラ	91	43
クラスタ番号９	しっかり自炊的	豆腐、卵、牛、豚、鳥肉、油、調味料	491	113
クラスタ番号１０	コープブランド的	全てコープブランド	119	15
クラスタ番号１１	健康飲料的	低脂肪牛乳、乳酸飲料	48	25
クラスタ番号１２	菓子のお伴的	ジュース、紅茶、	110	31
クラスタ番号１３	お手軽栄養的	バナナ、ヨーグルト、キウィ	150	36
クラスタ番号１４	肉不使用自炊的	やさい、魚、果物、	161	60
クラスタ番号１５	しっかり野菜的	トマト、白菜、玉ねぎ、にんじん	341	68
クラスタ番号１６	和風朝食的	豆腐、牛乳、ほうれん草、納豆、卵	173	34
クラスタ番号１７	おかずもう一品的	豆腐、みょうが、青シソ、ベーコン	207	68
クラスタ番号１８	見切り品的	野菜見切り品、果物見切り品	95	3
クラスタ番号１９	日用品的	生活用品催事、婦人用衣類	136	32
クラスタ番号２０	肉自炊的	牛、豚、鳥肉	412	111

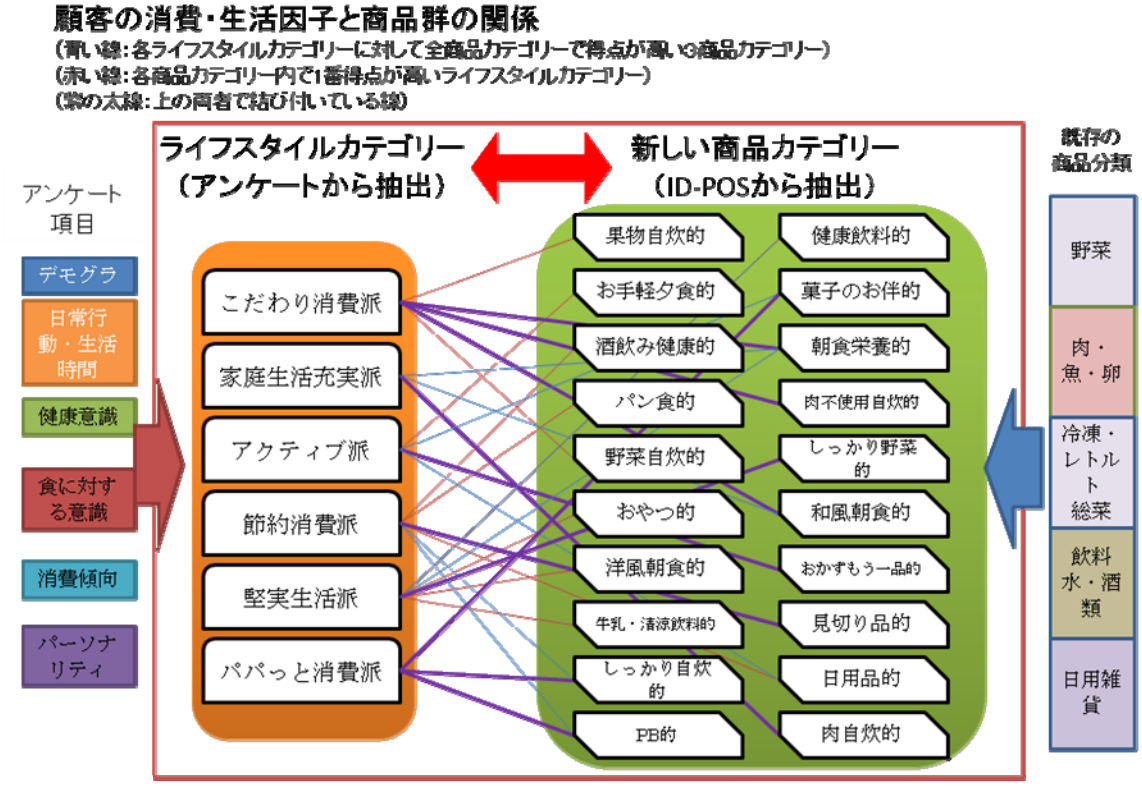

図 2.2.1-8　顧客のライフスタイルカテゴリと PLSI により分類された商品カテゴリとの関係

各商品カテゴリには同時に顧客も分類されている。そのため、顧客が持つ消費・生活因子得点を商品カテゴリ毎に計算することができる。図 2.2.1-9 にその一例を示す。図中の得点は、各商品カテゴリに属する顧客が持つ消費・生活因子得点の顧客数に対する平均値から、各因子得点に対する平均値の差をとったものである。つまり、ある因子の得点の値がゼロより大きければ、その商品カテゴリに属する顧客はその得点に対して強い回答をした傾向にあることを示している。ここでは各因子の得点の傾向を残すため、6 つの因子間の分散は正規化していない。この図ではライフスタイルカテゴリの第 4 因子（節約消費派）と商品カテゴリ 19 番の見切り品的が強く結び付いているなど、ある程度の妥当性を確認することができる。

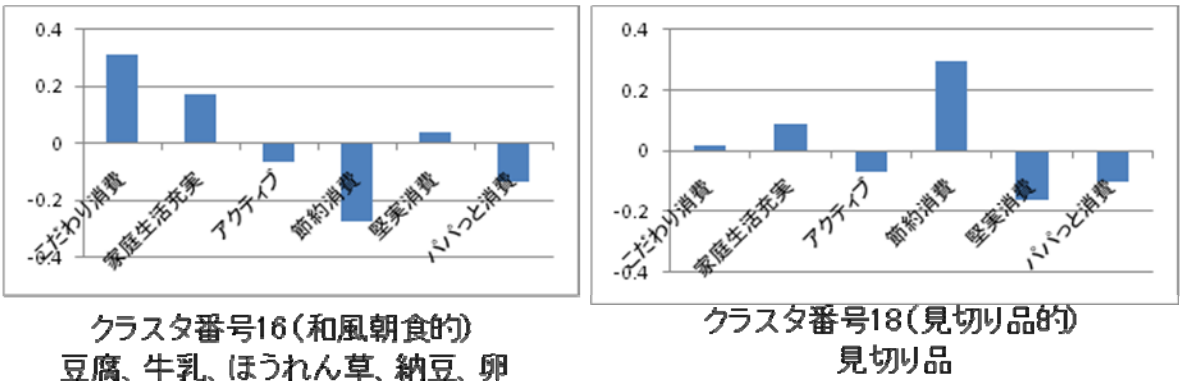

図 2.2.1-9　各商品カテゴリの消費・生活因子得点の例

表 2.2.1-4　密行列用 PLSI の実行時間とメモリ使用量

PLSIの実行時間評価(by Python2.6.4 64bit with Scipy Module)

	1000×1000行列						
				データ読み込み時間:0.7秒			
カテゴリ数	5	10	20	30	40	50	100
EMアルゴリズム1回の実測時間	0.6	1.2	2.3	3.7	5.1	6.1	12.9
PLSI処理の推定時間(EM×100反復)	1.0分	2.0分	4.3分	6.2分	8.5分	10.2分	21.6分
使用メモリ量(実測)	80MB	160MB	237MB	313MB	390MB	467MB	850MB

	10000×1000行列						
				データ読み込み時間:7.1秒			
カテゴリ数	5	10	20	30	40	50	100
EMアルゴリズム1回の実測時間[秒]	7.1	16.3	34.3	52.9	70.1	86.3	174.9
PLSI処理の推定時間(EM×100反復)	1.8分	2.7.2分	5.7.1分	8.8.2分	11.6.8分	2.4時間	4.9時間
使用メモリ量(実測)	972MB	1.3GB	2.1GB	2.8GB	3.5GB	4.5GB	8.0GB

	10000×10000行列						
				データ読み込み時間:63.3秒			
カテゴリ数	5	10	20	30	40	50	100
EMアルゴリズム1回の実測時間[秒]	90.4	190.0	405.3	600.2	実行不能	実行不能	実行不能
PLSI処理の推定時間(EM×100反復)	2.5時間	5.3時間	11.3時間	16.7時間	実行不能	実行不能	実行不能
使用メモリ量(実測)	6.0GB	10.3GB	17.7GB	26.1GB	実行不能	実行不能	実行不能

OS	Mac OS X バージョン10.6.1
プロセッサ	2×1.93GHz Quad-Core Intel Xeon
メモリ	32GB

各PLSI処理はカテゴリマイニングサーバを使用し、Python2.6.4の64bit版により動作させた。ここでは密共起行列に対応するため、行列の疎性を扱わないプログラムを作成した。また、計算速度を優先し、メモリ使用型のプログラム構成とした。その密行列用 PLSI の実行性能を表 2.2.1-4 に示す。また、Amazon 社が提供するクラウドコンピューティング（Amazon Elastic Mapreduce）を利用して PLSI を MapReduce フレームワーク内で並列分散処理するためのプログラムも開発した。しかしながら、現状のクラウドコンピューティングではデータ転送速度が本手法のために十分ではなく、1 GB のデータ転送に約 30 分必要となる。作成した MapReducePLSI プログラムでは 10000×10000 行列の計算に約 20GB のデータ量が必要となる。また、使用クラスタの指定やデータ配置などが取り扱えないため、冗長な動作が多く発生し、High-CPU ExtraLarge クラスタを 20 台使用しても 1 回の EM アルゴリズム反復に約 163 分かかってしまった。そのため、現状の Amazon EC2 の仕様では MapReduce フレームワークによる PLSI 処理は実用的ではないことが分かった。

一方、通常のデスクトップ PC を用いる実装であっても、1 年分、1 万人、1 万商品で 20 カテゴリの抽出にかかる速度は 11.3 時間、消費メモリは 17.7GB と十分実用になる結果である。本事業で開発した成果であるこのカテゴリマイニングソフトウェアとモデル構築ソフトウェアからなる大規模データモデル化技術に関しては３．１．４節 (2)において述べる。

(2-3) サイコグラフィックマイニング・分析作業のためのベイジアンネットモデル構築

ID-POS データ分析、アンケートデータ分析、PLSI による顧客分類の結果を用いてベイジアンネットモデルの構築を行った。ここでは、顧客の心理的特性（サイコグラフィック特性）に関するベイジアンネットモデルの構築を行うため、顧客特徴データベースを作成した。そのデータベースには以下の情報が含まれている。「顧客 ID、合計購入数、合計金額、購入平均単価、特定保健

用食品（トクホ）購入回数、プライベートブランド購入回数、国産品購入回数、健康食品購入回数、お手軽品購入回数、高級品購入回数、ダイエット的購入回数、お買い得購入回数、アンケート全 35 問、20 分類の所属商品カテゴリ」また、それらの情報は以下の条件で ID-POS データから抽出を行った。

・トクホ購入回数：特定保険商品を商品名で判断しラベル付け
・プライベートブランド購入回数：商品名に PB の名称が付いているものをラベル付け
・国産品購入回数：商品名に「国産」「＊＊県産」が付いているものをラベル付け
・健康食品購入回数：小分類が「291.健康志向茶」「625.健康食品」
・お手軽品購入回数：小分類が「21.野菜加工品」「60.刺身盛合せ」「65.フライ・ムニエル」「109.ローストビーフ」「112.牛肉加工品」「127.豚肉その他」「141.鶏肉タレ・味噌漬け」「142.鶏肉その他」「155.その他畜肉加工」「157.その他畜肉加工」「161.ミートデリカ」「162.鶏肉ミートデリカ」、中分類が「3.調理済み」「32.おかず」「33.スナック総菜」「34.総菜テナント」「35.冷凍食品」
・高級品購入回数：同小分類の商品の中で平均単価の比較的高いものをラベル付け（肉／魚など量が判定できないものは対象外）
・ダイエット的購入回数：商品名に「低脂肪」「低カロリー」「ダイエット」が付いているものをラベル付け
・お買い得購入回数：1 年間の平均単価と比較して、対象ジャーナル（買い上げ明細）1 レコードの税込金額/購入点数が 5%引きよりも安い場合にラベル付け

また、合計購入数、合計金額、購入平均単価、特定保健用食品（トクホ）購入回数、プライベートブランド購入回数、国産品購入回数、健康食品購入回数、お手軽品購入回数、高級品購入回数、ダイエット的購入回数、お買い得購入回数に関しては対象店舗の顧客に対して ABC 分析を行い、その結果をラベルとして付与した。

これらの顧客特徴量を用いて、ベイジアンネットモデル構築アプリケーション Bayonet を用いてベイジアンネットモデルの構築を行った。その確率構造モデルを図 2.2.1-10 に示す。その結果として、

・新商品が出ると試す人は、アクティブ消費派に分類される（男性より女性が多い）
・3 人以上の家族がいる家庭ではパパっと消費派になりやすい
・高くても健康重視の人はこだわり消費派に分類される
・産地レシピに興味がある人は家庭生活充実派になる傾向がある
・家計簿をつけている人は堅実生活派に分類される
・ドラッグストアの利用頻度が高いと節約消費派になる傾向がある

ジャーナルデータ（時間の入った買い上げ明細）に対し、上と同様の特徴量を付与しベイジアンネットモデルを構築した。その確率構造モデルを図 2.2.1-11 に示す。この図から、例えば下記のような傾向が示唆された。

- 家庭生活充実派の人は国産野菜を購入しやすく、お手軽品も購入しやすい
- 堅実生活派の人はお手軽品を購入しにくい
- お手軽品を購入している人は、お手軽夕食的・おやつ的・健康飲料的商品を購入しやすく、しっかり野菜的・果物自炊的商品は購入しにくい
- 堅実生活派の人は健康食品を購入しにくい
- 節約消費派の人はお買い得商品を購入しやすく、お買い得商品を購入している人はおかずもう一品的・見切り品的商品を購入しやすい。お手軽栄養的商品は購入しにくい
- アクティブ消費派はある PB 商品を購入しにくく、節約消費派の人は購入しやすい
- 節約消費派の人は、ダイエット的商品や洋風朝食的商品は購入しにくいが、しっかり自炊的商品は購入しやすい
- パパっと消費派の人は、菓子のお伴的商品やしっかり自炊的商品を購入しやすい

また、そのベイジアンネットモデルを用いて確率推論した一例を図 2.2.1-12 と図 2.2.1-13 に示す。図 2.2.1-12 はしっかり野菜的クラスタに含まれる商品の確率推論結果である。図中の紫色の線はしっかり野菜的クラスタの商品購買履歴の事前確率である。この商品群は午前中に多く購買されていることが分かる。また、こだわり消費派の顧客の購買率は約 5～9 ポイント低く、夏に購買率が上がる商品であることが分かる。特にこだわり消費派ではない顧客の夏の購買率は事前確率のポイントが 2 倍強となっていることが読み取れる。図 2.2.1-13 はお手軽品についての確率推論結果である。図中の紫色の線はお手軽品の商品購買履歴の事前確率である。この商品群は平日の夜に購買率が上がり、かつ夏の方がさらに購買率が上がることが読み取れる。

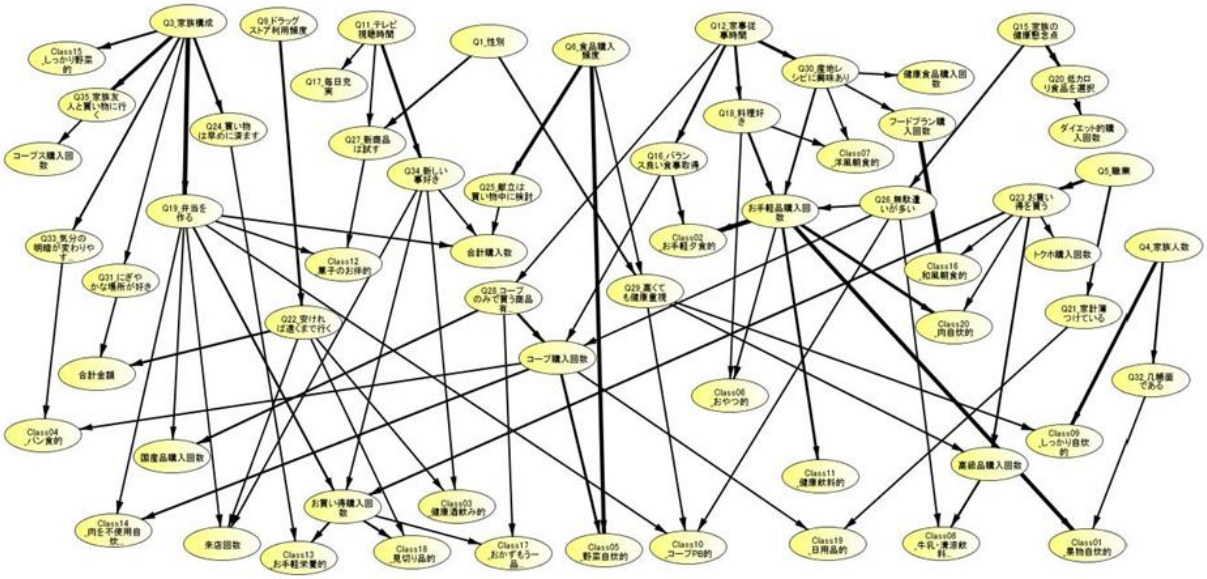

図 2.2.1-10　顧客特徴量から構築したベイジアンネットモデル

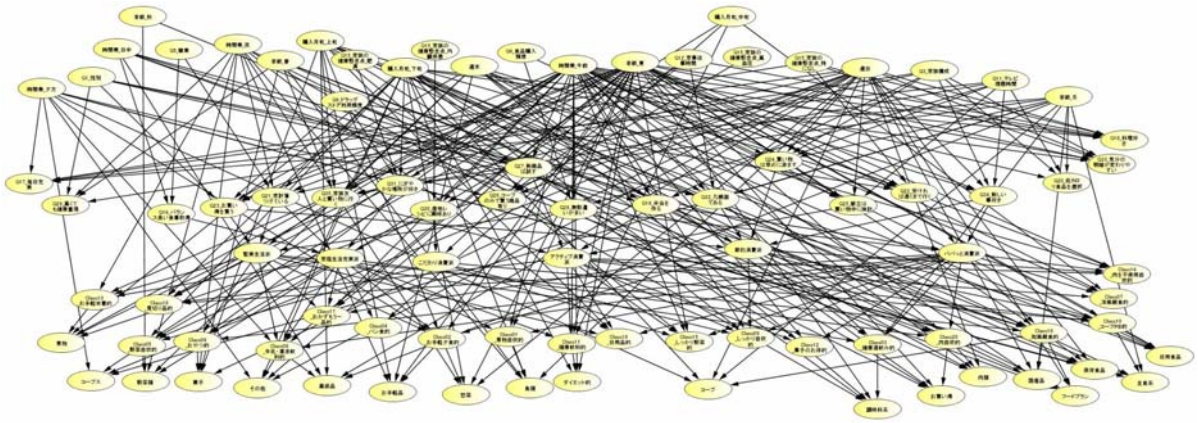

図 2.2.1-11　ジャーナルデータから構築したベイジアンネットモデル