ネットワーク分析に基づく顧客成長のための重要商品の抽出手法に関する一考察

情報数理応用研究　　5216C005-7　伊藤寛彬
指導教員　後藤正幸

A Study on Extraction of Important Items Focusing on Customer Growth Based on Network Analysis

ITO Hiroaki

1　研究背景・目的

近年の小売業では，会員システムやポイントシステムを利用し，顧客がいつ，どの店舗で，何を買ったのかについての詳細な購買履歴データを蓄積することが可能となった．このようなデータを活用したマーケティング分析も一般的になりつつあり，優良顧客をどのように獲得するのか，どのような施策を講じれば，離反顧客を抑止できるのかといった様々な観点からのデータ解析が数多く行われている．

本研究では共同研究先である株式会社良品計画の 1 年間の購買履歴データを対象事例として扱う．この小売企業における特徴の 1 つとして会員ステージ制度がある．この制度では，年間の累積購買金額に応じて 5 段階 (0〜4) の会員ステージが設定され，それぞれのステージに到達するごとに特典が設けられている．また，このステージは毎年 3 月 1 日に全ての顧客が初期ステージ (ステージ 0) から再スタートするという特徴がある．すなわちステージがリセットされる 2 月末の最終的な到達ステージは，優良顧客の度合いと捉えることもでき，多くの顧客に上位ステージまで到達してもらうことが望まれる．そのため，どのような商品を購買している顧客が優良顧客になっているのかについて分析することは重要と考えられる．

そこで，各ステージにおいて重要度の高い商品を特定し，上位ステージのみで重要度が高い商品を下位ステージの会員に購買してもらう施策を効果的に構成するために，重要な商品をステージ毎に特定する方法を検討する．こうした商品を下位ステージの会員に購入してもらうことで，より上位の会員ステージへの成長が期待できる．具体的には，各ステージの購買における商品間の共起の度合い (類似度) を用いて，商品をいくつかのクラスタへと分割する．さらに重要度分析によって，各ステージの重要商品や，ステージを上げるために重要な商品を抽出することで段階的な顧客成長を促進する方法を示す．

また，本研究では顧客の嗜好の多様性をよりネットワークに反映させるため，潜在クラスモデルの 1 つである Probablistic Latent Semantic Analysis と商品間遷移確率を導入して，ネットワーク分析と組み合わせた手法を提案する．以上により，本研究では構築した商品ネットワーク上において特定の顧客に購買されたことがある商品 (または，商品群) から，まだ購買されていない重要商品への商品推薦経路を考え，顧客に購買されたことがある商品とできるだけ特徴が似た商品を通りながら，徐々に重要商品へ近づけるような商品推薦の方法を示すことを目的とする．

2　準備

2.1　分析対象データ

本研究では，株式会社良品計画から提供頂いた無印良品の購買履歴データを分析対象とする．この小売ブランドでは食品や衣類，生活雑貨などの生活に必要なあらゆるカテゴリの商品を扱っている．購買履歴データの期間は 2014 年 3 月 1 日から 2015 年 2 月 28 日の 1 年間である．商品の購買は各店舗またはインターネット上の EC サイトで可能である．

本研究では，全顧客のうち，最頻利用店舗が東京都の店舗である 360,923 人を対象とし，購買商品についての分析をおこなう．ここで，EC サイトを含めた実店舗の総数は 447 店舗 (うち東京都にある店舗の総数は 83 店舗) である．また，商品ネットワークモデルを構築するための商品分類数は，当該店舗で販売されている商品の中分類 669 カテゴリを用いる．

2.2　基礎分析

事前分析により，2 月の最終時点でステージ 0 からステージ 4 であった顧客の購入商品を集計した結果，各ステージ間での売上個数や購買金額が多い商品に差異があることが明らかになった．そのため，商品推薦などの施策において会員ステージを上位まで引き上げるために重要度が高い商品も各ステージごとに異なると考えられる．

そこで最終的な到達ステージに着目し，着目したステージと上位ステージ間で，重要度が高い商品を抽出することができれば，顧客の成長を促す商品の特定が可能となり，マーケティング施策の立案に対する助けとなるものと思われる．

2.3　ネットワーク分析

本研究では，購買履歴データから得られる商品間のつながりを考慮した分析と，そのネットワークの可視化を目的とする．そこで商品間類似度を要素としてクラスタ分析と重要度分析を行う．ネットワーク構造におけるノードは各商品を，エッジは商品間類似度による商品間の繋がりを表す．商品間類似度が大きいほどエッジの繋がりが強いことを意味する．ここで，入力データとなる商品 a から商品 b への商品間類似度 (辺の重み) を s_{ab} とする．

本研究では，クラスタ分析と重要度分析の 2 つのネットワーク分析を行う．

2.3.1　クラスタ分析

ネットワーク構造のデータに対するクラスタ分析とは，与えられたグラフ構造から，ノードをつながりの密なグループへと分割する手法である．具体的には，グラフのリンク構造から，部分グラフ内のノード間のエッジ密度が，相対的に部分グラフ外への辺の密度よりも高くなるような部分グラフを切り出す．このことをクラスタリングまたはコミュニティ抽出という．

クラスタリングのための尺度としてはモジュラリティ Q を用いる．いま，抽出したコミュニティ集合を $\mathcal{C} = \{u, v, \cdots\}$，グラフ構造全体に含まれる辺の重みの合計数を m とする．ただし，コミュニティは 1 つ以上の商品集合から構成される．このとき，モジュラリティ Q は式 (1) で表現される．

$$Q = \sum_{u,v \in \mathcal{C}} \left(\frac{\sum_{a',b' \in u} s_{a'b'}}{2m} - \left(\frac{\sum_{a' \in u} \sum_{b' \in v} s_{a'b'}}{2m} \right)^2 \right) \quad (1)$$

このモジュラリティが高いほど，よいクラスタ分割ができていると解釈できる．モジュラリティ最大化は NP 困難であり，貪欲法を用いて高速に計算することが求められている．

2.3.2 重要度分析

ネットワーク分析における重要度分析とは，リンクのつながり方の重要度に応じてノードをランク化することである．これは，多くの重要なノードからリンクされているノードもまた重要であるという再帰的な考えに基づいている．N 個の商品をノードとしたとき，重要度の尺度として N 次元ベクトルである page rank $\boldsymbol{p}$ を用いる．この page rank $\boldsymbol{p}$ をノード毎に逐次更新し，得られた $\boldsymbol{p}^*$ が各ノードの重要度を示す．

ここで，A を正規化された隣接行列 (A_{ab}=$s_{ab}/s_{a\cdot}$) とする．なお，$s_{a\cdot}$ は s_{ab} の 2 番目の添え字 b を 1〜N まで動かして和をとったことを示している．r を所定のノード群にジャンプする確率，N 次元ベクトルである $\boldsymbol{q}$ を各ノードの重みとする．この重み $\boldsymbol{q}$ を用いることで各ノードに対し，顧客の嗜好に応じた重みを設定することができ，重みが高く設定されたノードにリンクした近傍ノードの重要度を高めることができる．page rank $\boldsymbol{p}$ は式 (2) で表現される．

$$\boldsymbol{p}^* = (1-r)A\boldsymbol{p} + r\boldsymbol{q} \tag{2}$$

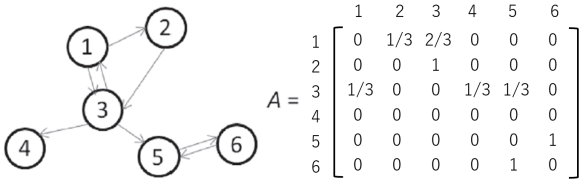

図 1．ネットワーク構造と隣接行列を用いた表現

2.4 PLSA

本研究では，購買データ分析に適した潜在クラスモデルとして PLSA(Probablistic Latent Semantic Analysis)[1] を適用する．PLSA は大規模な共起データから有用な知識を抽出する次元圧縮の手法であり，購買された商品と顧客の共起関係から，ソフトクラスタリングを行う．これにより，顧客嗜好の多様性や異質性を考慮しつつ，顧客の商品購買傾向をモデル化することができる．図 2 に PLSA のグラフィカルモデルを示す．

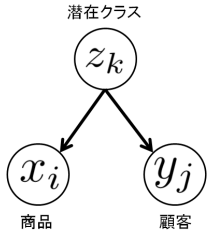

図 2．PLSA のグラフィカルモデル

いま I 個からなる商品集合を $\mathcal{X} = \{x_1, x_2, \cdots, x_I\}$，$J$ 人からなる顧客集合を $\mathcal{Y} = \{y_1, y_2, \cdots, y_J\}$，$K$ 個からなる潜在クラス集合を $\mathcal{Z} = \{z_1, z_2, \cdots, z_K\}$ とする．顧客 y_j が商品 x_i を購買する事象 (x_i,y_j) の確率は式 (3) で表される．

$$P(x_i, y_j) = \sum_{k=1}^{K} P(z_k)P(x_i|z_k)(y_j|z_k) \tag{3}$$

ここで，パラメータ $P(z_k)$，$P(x_i|z_k)$，$(y_j|z_k)$ は探索的手法の 1 つである EM アルゴリズム [2] により，式 (4) の対数尤度関数を局所的に最大化することで推定される．$\delta(x_i,y_j)$ は顧客 y_j が商品 x_i を購買しているかによって決定されるインジケータ関数である．

$$LL = \sum_{i=1}^{I}\sum_{j=1}^{J} \delta(x_i, y_j) \log P(x_i, y_j) \tag{4}$$

3 提案手法

3.1 概要

本研究では，2 つの分析手法を提案する．1 つ目は，単純な共起による商品間類似度を用いて，ステージごとの特徴を知るための，各ステージごとのクラスタ分析と重要度分析の 2 つを行う方法である．2 つ目は，顧客の購買嗜好に多様性があることや 2 商品間での遷移確率の違いがあることから，それらを考慮する手法を提案する．具体的には単純な共起による商品間類似度に対して，PLSA に基づく重み付けと，商品間遷移確率に基づく重み付けをする．この提案する商品間類似度を使用してネットワーク分析を行う．

3.2 単純な共起による商品間類似度を使用した各ステージごとのネットワーク分析

ステージごとに購買データを分け，全 2 商品間に対して，商品 a と商品 b を 1 年間で両方とも購買した顧客の人数に基づく 2 商品間類似度 s_{ab} を算出する．ただし，この類似度は対称，すなわち $s_{ab} = s_{ba}$ の関係があり，商品ネットワークを無向グラフとして考えたネットワーク分析を行う．また，商品 a と商品 b の組み合わせが購買された個数でなく，購買された人数とするのは，幅広い顧客に購買されやすい商品間のつながりを重視するためである．この商品間類似度 s_{ab} を用いて，各ステージごとの購買データに対してネットワーク分析によるクラスタ分析と重要度分析を適用する．

3.3 PLSA と商品間遷移確率を考慮した商品間類似度を使用した各ステージごとのネットワーク分析

3.2 節での提案手法では，商品間類似度 s_{ab} として，1 年間で 2 商品を両方とも購買した顧客の人数を用いた．本節では，顧客の購買嗜好を考慮した上で，嗜好が似た顧客に買われやすい商品をネットワーク上において近付けるために，PLSA に基づいた重みの算出を行うと共に，ネットワークを有向グラフに拡張するため，商品間遷移確率に基づいた重みの算出を行う．

これらの算出された重みをこれまでの単純な共起による商品間類似度 s_{ab} に掛け合わせることで，提案する商品間類似度 h_{ab} とし，ネットワーク分析を適用する．

3.3.1 PLSA に基づいた重みの算出

単純な共起による商品間類似度 s_{ab} を用いた場合，商品カテゴリや商品特徴が違う商品間での結びつきと，同じ商品間での結びつきをうまく考慮できない．事前分析をもとにすると，購買傾向は顧客によって非常に多様である．そのため，商品ネットワークを構築するにあたり，顧客の購買嗜好を考慮した上で嗜好が類似した顧客に購買されやすい商品同士の結びつきを強くしたほうが良いと考えられる．

そこで，購買商品と顧客の共起関係をソフトクラスタリングできる PLSA により，ユーザの購買嗜好および購買された商品の特徴をとらえることができると考えられる．ここで，商品 i が潜在クラス k に所属する確率を θ_{ik} としたとき，商品 i が各潜在クラスへの所属する確率を表す潜在クラス分布を $\boldsymbol{\theta}_i = (\theta_{i1}, \cdots, \theta_{iK})$ と表記する．商品 a の商品所属確率分布 $\boldsymbol{\theta}_a$ が商品 b の商品所属確率分布 $\boldsymbol{\theta}_b$ に対する分布間距離は式 (5) の Jensen-Shannon ダイバージェンス [3](以下，JS 情報量) を用いて求めることができる．

$$D_{JS}(\boldsymbol{\theta}_a||\boldsymbol{\theta}_b) = \frac{1}{2}(D_{KL}(\boldsymbol{\theta}_a||\boldsymbol{m}) + D_{KL}(\boldsymbol{\theta}_b||\boldsymbol{m})) \tag{5}$$

$$D_{KL}(\boldsymbol{\theta}_a||\boldsymbol{\theta}_b) = \sum_{k=1}^{K} \theta_{ak} \log \frac{\theta_{ak}}{\theta_{bk}} \tag{6}$$

$$\theta_{ik} = P(z_k|x_i) \tag{7}$$

$$\boldsymbol{m} = \frac{1}{2}(\boldsymbol{\theta}_a + \boldsymbol{\theta}_b) \tag{8}$$

$$w_{ab} = 1 - D_{JS}(\boldsymbol{\theta}_a||\boldsymbol{\theta}_b) \tag{9}$$

ここで式 (9) が示すように，商品所属分布が似た商品間ほどつながりが強くなるような重み w_{ab} を設定し，単純な共起による商品間類似度 s_{ab} に掛け合わせる．そうすることで購買商品と顧客の共起関係に基づく商品特徴が似た商品同士の類似度を高く設定できる．そのため，商品ネットワークを構築したときに嗜好が似た顧客に購買されやすい商品同士を近付けることができる．これにより，商品ネットワーク上における商品推薦のパスを考えたときに特徴が似た商品間を通りやすくなる．

3.3.2 商品間遷移確率に基づいた重みの算出

単純な共起による商品間類似度 s_{ab} を用いると，2 商品間を結ぶ 2 本のエッジの重みは双方とも等しいため無向グラフとなる．しかし，現実には 2 商品間における遷移確率には違いがあるため，それを考慮するためには有向グラフにする必要がある．

そこで，商品 a を購買したことがある顧客のうち，商品 b も購買したことがある顧客の割合 t_{ab} を算出する．この値を 2 商品間の重みと考え，単純な共起による商品間類似度 s_{ab} に掛け合わせる．そうすることでネットワーク上において商品 a から商品 b へ遷移する確率 t_{ab} と，商品 b から商品 a へ遷移する確率 t_{ba} の相違を表現できる．そのため，商品ネットワークを有向グラフとして構築することになる．

以上の 2 つの重みを考慮すると提案する商品間類似度 h_{ab} は式 (10) で表される．

$$h_{ab} = s_{ab} \times w_{ab} \times t_{ab} \tag{10}$$

最終的にこの構築された商品ネットワーク上において，ある商品（または商品群）を購買したことがある顧客に，重要商品を購買してもらうための最短経路問題を考える．このとき，単純な共起による商品間類似度 s_{ab} に比べ，PLSA と商品間遷移確率を考慮した商品間類似度 h_{ab} を使用したのときのほうが，商品ネットワーク構造において嗜好が類似した顧客に購買されやすい商品の推薦がされやすくなる．

4 分析

4.1 単純な共起による商品間類似度を使用した各ステージごとのネットワーク分析

単純な共起による商品間類似度 s_{ab} をもとに，各ステージごとに分割した購買データに対してクラスタ分析と重要度分析を行う．なお，クラスタ分析と重要度分析の具体的な演算には，NTT ソフトウェアイノベーションセンタが開発したグラフマイニングアプリケーションである Grapon[4] を用いた．

4.1.1 ステージごとのクラスタ分析

表 1 に各ステージごとの式 (1) のモジュラリティ値を示す．また，表 2 に例として，ステージ 0 とステージ 4 のクラスタ分析における各クラスの代表的な商品を示す．なお，ステージ 0 とステージ 4 のクラスタ分析において，最適なクラスタ数は双方とも 4 になった．

表 1 より，ステージが上がるに従ってモジュラリティ値が減少する傾向にある．このことは，ネットワーク上における商品のまとまりが複雑になっていることを表しており，優良顧客になるにつれて購買される商品が多様になっていくことが伺える．また，表 2 より各クラスの解釈をすると，ステージ 4(優良顧客）では食品・婦人服・雑貨と違うカテゴリのアイテムが混在したクラスが存在し，ここからも上位ステージにおいて，より多様な購買傾向があるということがわかる．

表 1. 各ステージにおけるモジュラリティ値

ステージ	0	1	2	3	4
モジュラリティ値	0.1303	0.0870	0.0829	0.0732	0.0769

表 2. ステージ 0 とステージ 4 における各クラスの解釈

クラス	ステージ 0	ステージ 4
1	衣類	紳士服
2	食品	食品，婦人服，雑貨
3	子供用商品	子供用商品
4	生活雑貨	生活雑貨

4.1.2 ステージごとの重要度分析

重要度分析において式 (2) の重み $\boldsymbol{q}$ は，全ての商品に関して等しいとして 1 に，所定のノードにジャンプする確率 r は一般的によく用いられる 0.8 に設定する．また，重要度計算を高速化するために，計算過程で重要度が低いノードの枝刈りをし，最終的に重要度が高い上位 100 個のノードのみを抽出する．表 3 に各ステージの重要商品上位 5 商品を示す．ステージ 0 からステージが上がるにつれて重要度が上がっていく商品はステージ 4 の上位 20 位の中に 9 商品あり，その全てが生活雑貨である．表 3 からもステージが上がると主に生活雑貨の重要度の順位が上昇する傾向にあることがわかる．反対にステージが低いときは食品の重要度が高く，食品が無印良品ブランドで購買するきっかけになりやすいと考えられる．また，上位ステージほど商品重要度の偏りが少なく，幅広く商品を購買する傾向にあり，下位ステージほど商品重要度が上位の商品に偏っており，特定の商品に依存していることがわかる．

表 3. 各ステージの重要度上位アイテム

ステージ 0		ステージ 1		ステージ 2	
商品名	重要度	商品名	重要度	商品名	重要度
食品 A	0.01004	食品 A	0.00816	食品 A	0.00702
食品 B	0.00900	生活雑貨 A	0.00736	生活雑貨 B	0.00653
生活雑貨 A	0.00880	食品 B	0.00731	生活雑貨 A	0.00635
食品 C	0.00837	生活雑貨 B	0.00718	食品 B	0.00630
生活雑貨 B	0.00795	食品 C	0.00682	食品 C	0.00592

ステージ 3		ステージ 4	
商品名	重要度	商品名	重要度
食品 A	0.00606	食品 A	0.00474
生活雑貨 B	0.00578	生活雑貨 B	0.00471
食品 B	0.00552	生活雑貨 J	0.00459
生活雑貨 A	0.00544	生活雑貨 M	0.00459
食品 C	0.00514	生活雑貨 K	0.00459

4.2 PLSA と商品間遷移確率を考慮した商品間類似度を使用した各ステージごとのネットワーク分析

4.1 節と同様の方法で，PLSA と商品間遷移確率を考慮した商品間類似度 h_{ab} をもとに，各ステージごとにネットワーク分析を行う．

例としてステージ 0 における商品間類似度 s_{ab}，h_{ab} を用いたときの商品ネットワークをそれぞれ図 3，図 4 に示す．ここでは，グラフ可視化アプリケーションである Gephi[5] を使用して可視化を行う．図においてノードの大きさが商品の重要度を，矢印の太さが商品間のつながりの強さを表している．商品間類似度 s_{ab} を用いた商品ネットワーク (図 3) では重要商品間でのつながりが強く，商品のまとまりが悪い．一方で商品間類似度 h_{ab} を用いた商品ネットワーク (図 4) では，ネットワークの中央に

重要商品がまとまり，その周辺にカテゴリごとの商品がまとまる．すなわち，単に商品間の共起情報だけでなく，PLSA による各潜在クラスへの所属度や商品間遷移確率を考慮することでより顧客の嗜好を考慮したネットワークの構築ができたといえる．さらに，図 4 では重要商品と各カテゴリをつなぐような橋渡し商品も確認できる．

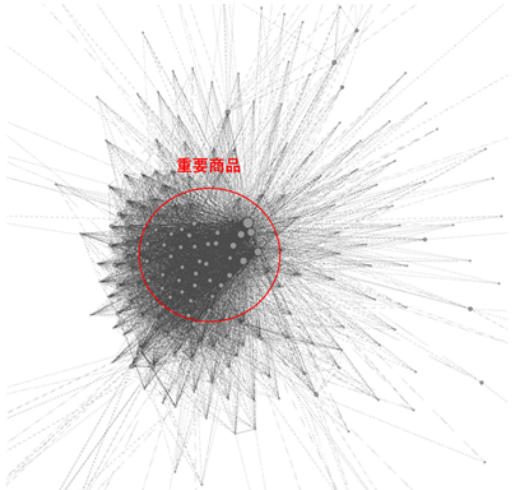

図 3．商品間類似度 s_{ab} を使用した
ステージ 0 の商品ネットワーク

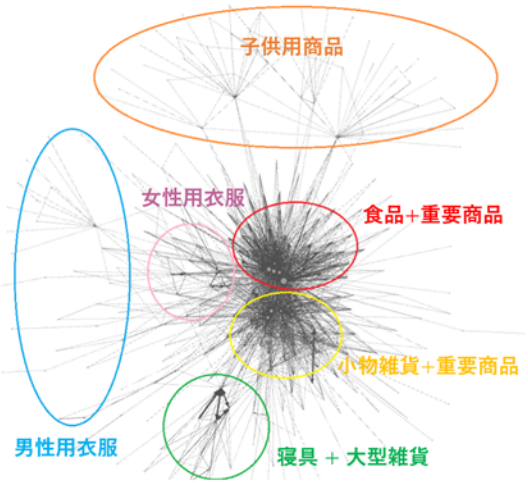

図 4．商品間類似度 h_{ab} を使用した
ステージ 0 の商品ネットワーク

表 4 にステージ 0 とステージ 4 における商品間類似度 s_{ab}，h_{ab} を用いた場合の式 (1) のモジュラリティ値を示す．両ステージにおいて商品間類似度 h_{ab} を用いることでモジュラリティ値が向上しており，嗜好が似た顧客に購買されやすい商品同士をネットワーク上においてまとめることができたためであると推測される．

表 4. 従来と提案でのモジュラリティ値の比較

ステージ	商品間類似度 s_{ab} を使用	商品間類似度 h_{ab} を使用
ステージ 0	0.1303	0.4078
ステージ 4	0.0769	0.2038

表 5 に商品間類似度 h_{ab} を用いたときの重要度上位商品を示す．商品間類似度 s_{ab} を用いたときに上位に出現した商品は，同じく上位に出現している．しかし，商品間類似度 h_{ab} のときの上位商品の中には商品間類似度 s_{ab} のときに下位にあった商品も多く含まれることがわかる．これらの商品の多くは図 4 における重要商品と各カテゴリの商品を橋渡しするような商品であった．

最終的に構築した商品ネットワーク上においての最短経路問題を解くことで，商品推薦経路を決定する．商品間類似度 h_{ab} を用いることで，特徴が類似した商品間を推薦しやすくなるため，店舗のレイアウトや顧客嗜好の点で現実的な商品推薦が可能となる．

表 5. 商品間類似度 h_{ab} を使用したときの
ステージ 0 の重要商品上位

順位	商品	重要度	従来の順位
1	食品 A	0.0465	1
2	食品 B	0.0287	2
3	食品 C	0.0238	4
4	生活雑貨 A	0.0212	3
5	食品 D	0.0183	6
6	生活雑貨 B	0.0181	5
7	衣類 D	0.0153	**57**
8	衣類 E	0.0122	**99**
9	生活雑貨 D	0.0121	8
10	生活雑貨 C	0.0112	7
11	衣類 F	0.0102	**86**
12	生活雑貨 E	0.009	11
13	生活雑貨 P	0.0089	**27**
14	食品 F	0.0075	10
15	食品 E	0.0072	9
16	衣類 G	0.0071	**100 位以下**
17	食品 G	0.0068	14
18	生活雑貨 W	0.0064	**37**
19	衣類 H	0.0063	**32**
20	生活雑貨 G	0.0063	13

5　まとめと今後の課題

本研究ではまず，2 つの商品を 1 年間で購買した人数をこれらの商品間の類似度としてグラフを生成し，ステージごとにクラスタ分析と重要度分析を行った．

さらに潜在クラスモデルである PLSA に基づいた商品の類似性を示す重みと，商品間遷移確率に基づく重みを算出する．これらの重みを単純な共起による商品間類似度と掛け合わせることで商品ネットワークを有向グラフと考え，クラスタ分析と重要度分析を適用し，その可視化を行った．これにより，嗜好が類似した顧客に購買されやすい商品を推薦しやすくなるため，店舗のレイアウトや顧客嗜好を考慮したプロモーション施策などの点においてより現実的な商品推薦を可能とした．

今後の課題としては，PLSA における潜在クラス数の最適な決定が挙げられる．また本研究では，商品のみのネットワークを構築したが，商品に加え顧客も取り入れたネットワークを構築することで，商品と顧客または顧客と顧客のつながりを把握できるようにすることも今後の課題として挙げられる．

参考文献

[1] Hofmann，T.，“Probabilistic latent semantic indexing,” *Proc. 22nd Annusl ACM Conference on Research and Development in Information Retrieval*, pp. 50-57, 1999.

[2] Dempster, A. P., Laird, N. M., Rubin, D. B., “Maximum likelihood from incomplete data via the EM algorithm,” *J. Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1-38, 1977.

[3] Fuglede, B., Topsoe, F., “Jensen-shannon divergence and hilbert space embedding,” *Proc. Internationales Symposium on Information Theory*, pp. 31-39, 2004.

[4] 飯田恭弘，岸本康成，藤原靖宏，塩川浩昭，鬼塚真，“大規模グラフ構造データからのコミュニティ抽出と重要度計算— 高速化への取組みと応用—,” 人工知能，vol. 29, no. 5, pp. 472-479, 2014.

[5] Bastian, M., Heymann, S. and Jacomy, M., “Gephi : An open source software for exploring and manipulating networks,” *Proc. ICWSM*, 2009.