論文・事例研究

グラフラプラシアンを用いたCDの特徴抽出とその利用

大野　尚志，矢島　安敏

1. はじめに

本研究では，「平成17年度データ解析コンペティション」において提供されたCD販売店の購買履歴データを使い，顧客の購買傾向の分析を行う．また，それに基づきCDの特徴抽出を行い，顧客へのCDのリコメンドおよび売上パターン予測を試みる．

分析に用いたデータには，CD販売店10店舗分(内，少なくとも1店舗は都内）のCD購買履歴データが含まれている．本研究では，店舗の区別はせず10店舗全てのデータについて，2003年9月から2005年8月までの2年間分を用いた．購買履歴の各レコードは，1回の購買毎に，顧客ID，利用年月日，商品IDから構成されている．ここで商品IDとは，商品毎に付与されたJANコードのことで，このコードを基に各CDの商品名やアーティスト名，ジャンル名といった情報を得ることができる．

本研究で用いたデータにおける特徴として，CD1枚あたりの売上枚数の少なさがある．図1は，データ中における各CD1枚あたりの売上枚数を集計し，ヒストグラムにしたものである（ただし，売上枚数100枚以上のCDは省略した)．この図からもあきらかなように，大部分のCDが売上枚数わずか数十枚という状況である．実際，全CDの売上枚数の平均は32枚であり，売上枚数53枚以下のCDが全CDの約90%以上を占めており，購買履歴中に約3万種類のCDが記録されているのと比べて極めて少ない．また，異なる2枚のCDに関して両方ともに購入している顧客数を調べてみると，多くのCDの組合せで「0人」であり，数人以下の組合せがほとんど全てである．すなわち，CD間の関連性等を購入した顧客の購買データで

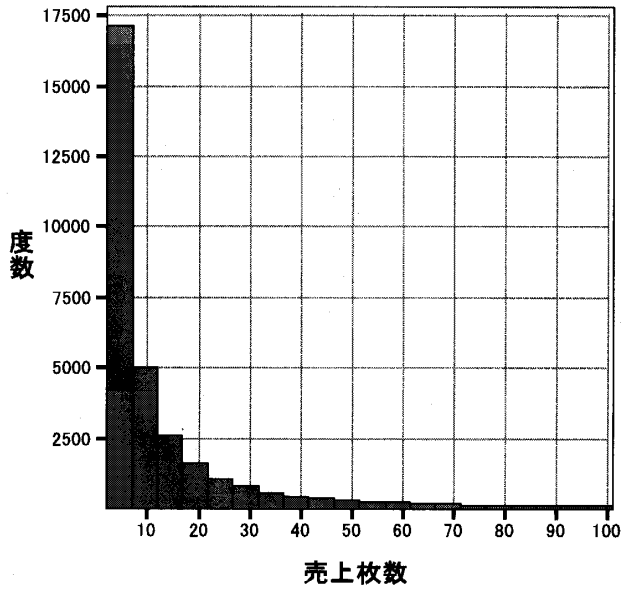

図1　売上枚数の分布

説明しようとしても，このままでは関連性を適切に定めることが困難である．本研究では，購買履歴データをグラフとして表現することにより，このような履歴データの少なさにも対応した分析を試みる．

次の2節では，まず，グラフラプラシアンに基づき，CD間にある種の類似性を定める手法について述べる．また，2クラスの判別法であるサポートベクターマシン（Support Vector Machine, SVM）に代表される，カーネル法との関連性についても述べる．さらに，得られた類似性に基づきCDの関連性を視覚化し，類似性の妥当性について検討を行う．続く3節では，実際に顧客へのCDのリコメンド（購買顧客予測）およびCDの売上パターン予測を試み，その精度を検証した．最後に4節でまとめを述べる．

2. グラフ構造を用いたCDの特徴抽出

上で述べたように，本研究で扱うデータでは実際に購買されている商品が上位のもののみに集中しており，顧客の購買行動（購買パターン）からCDや顧客の特

おおの　なおし，やじま　やすとし
東京工業大学 大学院社会理工学研究科
〒152-8552 目黒区大岡山2-12-1
受付06.7.21　採択06.11.15

徴を分析することは困難である．このようなデータに対して，近年isomap[9]やLLE[7]，ラプラシアンカーネル[1]などの「局所的」な類似性に基づいたパターン認識の手法が提案され，電子メールなどのテキスト文書やWebページの分類，あるいは画像の自動識別といった分野で盛んに用いられている．これらの手法は，類似性の非常に高い部分の情報（局所的な類似性）のみを用いてグラフを構築し，その上でグラフの全体的な構造によりデータの類似性あるいは距離を定める方法である．この節では，本研究で用いるラプラシアンカーネルによる方法について説明する．

2.1 グラフを用いたデータの表現

まず，与えられたデータ全体を重み付きグラフ$G(V,E)$を用いて表現する．Vはノード集合で，各ノードは「データの要素（data item）」に対応している．data itemとはデータを構成している個々の要素のことで，例えば，上であげたパターン認識の応用例の場合では，個々のテキスト文書やWebページ，あるいは画像のことを指し示す．本研究で扱うデータの場合では，個々の顧客およびCDである．以降では，特に混乱が無い限りdata itemを単にデータと呼ぶこととする．Eは枝の集合で，ノード$i,j\in V$間に枝$(i,j)\in E$が存在すれば，ノード間の類似性の大きさを表す正の重みb_{ij}を付与する．便宜上，枝(i,j)が無い場合には重さを$b_{ij}=0$と考える．

枝の重みの与え方にはさまざまな方法が知られている．中でも，類似性の高いノード間にのみ枝を与え，全ての枝(i,j)に単純に$b_{ij}=1$と重み付けをしたグラフがよく用いられる．例えば，各ノードに対して，他の全てのノードの中からもっとも類似性の高いk個のノードを選び出し，それらに対してのみ枝を張ったk最近傍グラフなどが用いられる．本研究の場合では，注目したいデータは顧客とCDである．そこで，顧客とCDをそれぞれノードとしたグラフを考える．また，購買履歴より，顧客がCDを購入していれば，局所的な類似性があると考え，対応する顧客とCDのノード間に枝を張ったグラフを構築した．すなわち，二部グラフが作成される．枝の重みは単純に全て$b_{ij}=1$と設定した．

以降では，議論を簡単にするために，構成されたグラフは連結であることを仮定する．また，ノード数をMとして，重みb_{ij}を$i-j$成分とするM次の対称行列をBと記す．行列Bはその定義よりデータ間の類似性の強さを表す行列であるが，一般に0の要素が多い非常にスパースな行列であること，さらに，要素が0の部分は，必ずしも類似性が0であることを意味しない点に注意して欲しい．Bの非ゼロ部分は「局所的な類似性」を表しており，これを基にしてグラフの全体的な構造を反映させることで全てのノード間に適切な類似性を定める方法のひとつに，グラフラプラシアンを用いた方法がある．以降，これについて説明する．

2.2 離散時間マルコフ連鎖モデル

まず，各ノード$i\in N$での重みb_{ij}の和$D_{ii}=\sum_{j=1}^{M}b_{ij}$を対角成分とした対角行列を$D$とする．また，$I$を$M$次の単位行列として，$0<\alpha<1$をパラメータとした次の行列

$$P=(1-\alpha)I+\alpha D^{-1}B$$

を定める．行列Pの$i-j$要素をp_{ij}と記せば，あきらかにp_{ij}は非負の値で，かつ各行の要素の和は$\sum_{j=1}^{M}p_{ij}=1$となることより，Pは推移確率行列と考えることができる．そこで，グラフ上をノードiからjへと確率p_{ij}で推移する離散時間マルコフ連鎖を考える．

今，ノードiから出発したランダムウォーカーが初めてノードjへ至るfirst passage timeの期待値をh_{ij}とし，さらに，次のような和

$$n_{ij}=h_{ij}+h_{ji}$$

を定める．n_{ij}はiからjに到達し再びiへ戻るまでの期待時間であり，commute timeと呼ばれている．あきらかにn_{ij}は非負の値で対称（$n_{ij}=n_{ji}$）である．また，$\sqrt{n_{ij}}$は三角不等式を満たすことが知られている．そこでラプラシアンカーネルを用いた手法では，n_{ij}をノード間の距離（の2乗），すなわち**非類似性**と考える．n_{ij}は離散時間マルコフ連鎖の性質を考えると，

(i) ノードi，jが離れているほど大きな値となる

(ii) ノードi，j間に多くの種類のパスがあるほど小さな値となる

といった性質を持つこととなり，グラフ全体の枝の接続の様子を反映したものとなる．また，$G=(D-\alpha B)^{-1}$とし，行列Gの$i-j$要素をg_{ij}とすれば，

$$n_{ij}\propto g_{ii}+g_{jj}-2g_{ij}$$

と求められることが知られている[4]．

さらに，次のようにある種の基準化を施したnormalized commute timeを用いることで，より優れた結果を導くことが経験的に知られており，SVMをはじめとしたカーネルを用いたパターン分類などに用い

られ，高い判別力を示すことが知られている[10]．まず，$\pi=(\pi_1, \pi_2, ..., \pi_M)$ を定常分布を示す確率，すなわち $\pi=\pi P$ を満たす行ベクトルとする．π を使い first passage time の期待値を $\bar{h}_{ij}=\sqrt{\pi_i \pi_j} h_{ij}$ と基準化し，normalized commute time を

$$\bar{n}_{ij}=\bar{h}_{ij}+\bar{h}_{ji}$$

と定める．

ここで，$\pi_i=D_{ii}/\sum_{k=1}^{M} D_{kk} (i=1, 2, ..., M)$ であることを使えば，行列

$$\bar{G}=(I-\alpha D^{-1/2} B D^{-1/2})^{-1} \quad (1)$$

の要素を $\bar{g}_{ij}$ とすれば，normalized commute time は

$$\bar{n}_{ij} \propto \bar{g}_{ii}+\bar{g}_{jj}-2\bar{g}_{ij} \quad (2)$$

となることが知られている．

2.3 カーネル法との関連性

最後に，式(1)で定めた行列 $\bar{G}$ と，SVM などに用いられるカーネル行列との関連性について述べる．まず，グラフラプラシアン[3]と呼ばれる次の行列

$$L=I-D^{-1/2} B D^{-1/2}$$

を定める．行列 L は半正定値行列となることが知られている．さらに，σ を実数のパラメータとして次の行列：

$$K_L=(I+\sigma^2 L)^{-1}$$

を定めれば，$I+\sigma^2 L$ が正定値行列であることより，K_L も正定値行列となり，これはラプラシアンカーネル[8]と呼ばれる行列である．近年，このカーネル行列を判別問題に応用することで，高い性能を発揮することが知られている[10]．簡単な式変形で，

$$K_L=\frac{1}{1+\sigma^2}\left(I-\frac{\sigma^2}{1+\sigma^2} D^{-1/2} B D^{-1/2}\right)^{-1}$$

$$=\frac{1}{1+\sigma^2}\bar{G}$$

となることより，行列 $\bar{G}$ はラプラシアンカーネルと考えることができる．

さて，SVM などに代表されるカーネル法は，データとデータとの内積の値を定めることで分析を行う方法である．通常のデータ分析であれば，まずデータをいくつかの属性によって特徴づけ，属性の値を要素とするベクトルとして表現することが必要となるが，これは以下に述べるように内積の値を定めることとほぼ同値な作業である．カーネル法では，例えば，上で導入したラプラシアンカーネル行列 K_L の $i-j$ 成分 k_{ij} はグラフのノード i，j にそれぞれ対応したデータの内積である．カーネル行列 K_L は半正定値行列であることから，U を K_L の固有ベクトルを列ベクトルとする正規直交行列，Λ をその固有値を対角要素とする対角行列とすれば，Λ の対角要素が非負であることから $K_L=U\Lambda U^T=(U\Lambda^{1/2})(U\Lambda^{1/2})^T$ と分解可能である．そこで，行列 $\Lambda^{1/2} U^T$ の列ベクトルを

$$\Lambda^{1/2} U^T=[\phi_1 \cdots \phi_l] \quad (3)$$

とすれば，ϕ_i はノード i に対応したデータをベクトルとして表現したものと考えることができ，ϕ_i と ϕ_j の内積は，

$$k_{ij}=\phi_i^T \phi_j, \ i, j=1, 2, ..., l$$

と K_L の $i-j$ 要素と一致する．すなわち，データ間の内積を半正定値なカーネル行列で与えることは，データをベクトルとして表現することと等価である．

以上の議論を用いれば，データ i，j 間のユークリッド距離（非類似度）の2乗は，

$$\|\phi_i-\phi_j\|^2=k_{ii}+k_{jj}-2k_{ij} \quad (4)$$

とカーネル行列により与えられる．さらに式(2)を見ればあきらかなように，ラプラシアンカーネル K_L が定める距離は，normalized commute time であることが分かる．

2.4 CD の特徴の可視化

さて，上で述べたように，カーネル行列は内積を要素とするものであることから，カーネル行列 K_L を用いて多次元尺度構成法（MDS）[2]を実行すれば，ベクトル ϕ_i を低い次元の空間に布置し，その特徴を視覚的に表現することが可能である．これには，K_L の固有値と固有ベクトルが必要で，それぞれ，$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_M \geq 0$ および $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_M \in \mathbf{R}^M$ とする．MDS で注目するのは，大きな固有値に対応した固有ベクトルである．例えば，大きな固有値に対応した2本の固有ベクトル $\boldsymbol{u}_1$，$\boldsymbol{u}_2$ を用いた場合，それぞれのベクトルの j 番目の要素を u_{j1}，u_{j2} とすれば，データ j の2次元平面上の座標は $(\sqrt{\mu_1} u_{j1}, \sqrt{\mu_2} u_{j2})$ と表される．

本研究の場合，カーネル行列 K_L の次元は顧客数と CD 数の和となり極めて大きい．一般に，このような巨大なサイズの行列の逆行列を扱うことは困難である．しかし，ラプラシアンカーネルの場合，$K_L=(I+\sigma^2 L)^{-1}$ となっていることより，K_L の固有ベクトル $\boldsymbol{u}_i$ は，グラフラプラシアン L の固有ベクトルに他ならない．また，K_L は逆行列として与えられていることから，L の小さな固有値に対応する固有ベクトルを求めれば，K_L の大きな固有値に対応したものとなる．さらに，本研究の場合，L は2部グラフの隣接行列に相当するため極めてスパースな行列となる．そ

こで，Lanczos法などを用いることで，10万次元を超えるような大規模な行列であっても容易に固有ベクトルを算出することが可能である．さらに，$\boldsymbol{e}\in\mathbf{R}^M$ を要素が全て1のベクトルとすれば，$L(D^{\frac{1}{2}}\boldsymbol{e})=0$ となることより，L の最小固有値は0，また，これに対応する固有ベクトルは $\boldsymbol{u}_1=D^{\frac{1}{2}}\boldsymbol{e}$ である．すなわち，座標の値は顧客であれば購入したCDの数，CDであれば購入した顧客の人数に相当する．

そこで，$\boldsymbol{u}_1$ 以外に，さらに4本の固有ベクトル $\boldsymbol{u}_2$ から $\boldsymbol{u}_5$ を算出し，売上総数100枚以上のCD約2,000枚について布置したものが図2，3である．$\boldsymbol{u}_2$，$\boldsymbol{u}_3$ の2本を用いて布置したものが図2，また，$\boldsymbol{u}_4$，$\boldsymbol{u}_5$ を用いたものが図3であり，これら2つの図は直交した関係にある．なお，図中には特徴的なCDについて，

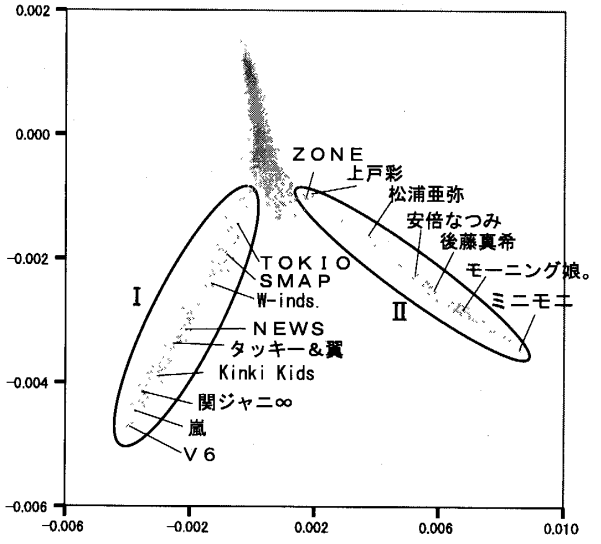

図2　CDの特徴の可視化の例(1)

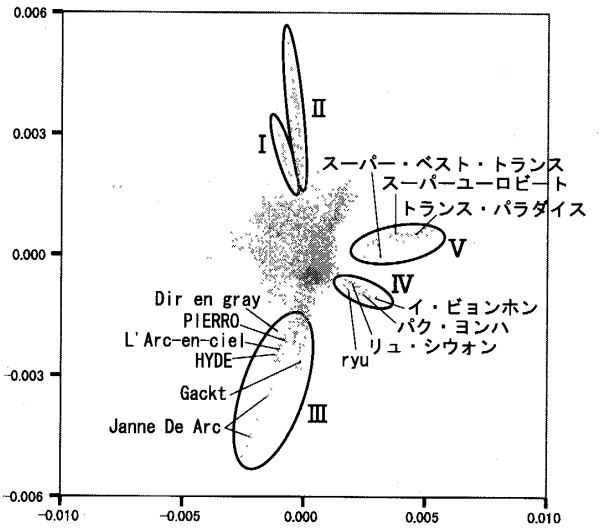

図3　CDの特徴の可視化の例(2)

そのCDの歌手名またはCDタイトルを付与した．

まず，図2を見てみると，2つの特徴的なグループができていることがわかる．ここで，各グループのCDの歌手名を見てみると，グループIは男性アイドル歌手，そしてグループIIは女性アイドル歌手のグループであることがわかる．このように，異なる場所に布置されるグループは，その購入者層が非常に異なることを示唆している．また，V6や嵐，ミニモニなどのアイドル歌手のCDは他のCDとは離れた場所に布置されていることから，同様な歌手を好む一部の熱狂的なファンによって購入される傾向があると考えられる．一方，ZONE，上戸彩，松浦亜弥，TOKIO，SMAPなどは比較的一般的な顧客に支持されている歌手であると思われる．

次に図3を見ると，図2で示されたグループI，IIがほぼ1つに重なるようにまとまり，それに代わり図2では重なってしまい見ることのできなかった特徴的なグループを得ることができる．これらのCDのタイトルから，グループIIIはヴィジュアル系アーティストのグループ，グループIVは韓国人歌手のグループ，そしてグループVはクラブ・ダンス系のグループである．I，II同様，これらのグループも他とは異なる購買層を持っていることを示唆していると考えられる．

次に複数枚のシングルCDをリリースしている歌手について，歌手毎にCD間の距離を式(4)により算出し，それらの最大値を歌手のバラエティの豊かさと考えてみた．すなわち，距離の離れたCDをリリースしている歌手ほど，幅広いジャンルの曲を歌いこなせる歌唱力のある歌手と考えられる．ここでは，シングルCDの売上枚数1,000枚以上の歌手66名のCD約1,000枚について，内積 k_{ij} を求めた．カーネル行列 K_L 全体を $(I+\sigma^2L)$ の逆行列により求めることは，その大きさから不可能であるが，あるCDに対応する K_L の列はCG法などで算出が可能である．すなわち，$\boldsymbol{e}_i$ を単位行列 I の第 i 列ベクトルとすれば，K_L の第 i 列は方程式 $(I+\sigma^2L)\boldsymbol{x}=\boldsymbol{e}_i$ により求めることが可能である．ちなみに，約1,000枚のCDについて計算をしても，Pentium 4 (3.8 GHz) のパソコンで数十分程度の計算時間であった．表1には，代表的な歌手名を示した．CD間の距離の離れている歌手（グループ）としてはサザンオールスターズなどが上位になっており，異なるさまざまな曲を歌っていることが分かる．また，CD間の距離の近いものとして挙げられている歌手は，どれを聞いても比較的同じような曲を歌

表1　CD間距離による歌手の分類

CD間距離の離れている代表的な歌手
サザンオールスターズ, 森山直太朗, 平原綾香, 福山雅治, ケツメイシ, くるり, 倖田來未, BUMP OF CHICKEN, オレンジレンジ
CD間距離の近い代表的な歌手
aiko, V6, ゆず, GLAY, TOKIO, ASIAN KUNG−FU GENERATION, 玉置成実, L'Arc〜en〜Ciel, モーニング娘。, 松浦亜弥

っている歌手である．

本研究では，購買の有無のみを考え枝を与えた二部グラフを用い，また重みも一律に1とした単純なグラフを構築したが，視覚化を行った結果，性質の似ていると思われるCD同士は互いに近い場所に布置され，また，比較的購買層に偏りがあると思われるCDは他のCDから離れた場所に布置される結果を得た．このことから，構築したグラフと離散時間マルコフ連鎖モデルに基づいたnormalized commute timeを用いた距離は，顧客の購買行動やCDの特徴を正しく反映していると考えることができる．

3. 抽出したCDの特徴の利用

前節で導入したラプラシアンカーネルを用い，顧客へのCDのリコメンド（顧客の購買予測）およびCDの売上パターン予測を行い，その精度を検証する．

3.1 顧客へのCDのリコメンド

まず，2年間の総売上枚数が2枚以上のCD約3万タイトルを選び，これらのCDを2種類以上購入した顧客約18万人を対象としてリコメンドを行った．2004年1月1日から同年8月31日までの間に発売されたCDの中で，売上枚数上位20タイトルそれぞれについて，上で述べたように二部グラフを構築し，ラプラシアンカーネルによりCDと顧客との類似性を算出し，以下に述べる方法により，手法の精度を検証した．

提供されたデータには，購買が行われた日時が記録されている．そこで，CDの発売当日までのデータに限定してグラフを構築し，そのCDを未購入の顧客に対して類似性を計算し，購買可能性を表す指標とする．すなわち，CDの発売初日1日分およびそれ以前の全ての購買履歴を学習用データとして用い，将来購買するであろう顧客の予測を行った．当然のことであるが，CDの発売日が異なれば構築されるグラフも異なる．さらに，手法の性質上，グラフ上で予測するCDに対応するノードと，連結なコンポーネントに属している顧客のみが予測が可能な顧客である．また，グラフを構築する時点で何も購入履歴の無い顧客に対しては予測ができず，この点は手法の限界のひとつである．なお，ラプラシアンカーネルのパラメータは$\sigma^2=10$を用いた．

表2　20作品の平均再現率（%）の比較

r	1%	2%	5%	10%	20%
提案手法	7.9	12.5	22.6	33.4	47.9
相関係数法	4.3	6.7	14.2	25.3	42.3

CDの発売日から1年後までにそのCDを実際に購買した顧客を，どれだけ正確に抽出できるかを再現率で表し，手法の精度を評価する．例えば，購買可能性上位r%での再現率とは，

$$\frac{\text{上位}\ r\%\text{顧客のうちCDを購入した顧客数}}{\text{CDを購入した顧客数}}$$

で計算される指標である．なお，比較のため，協調フィルタリングの代表的手法である相関係数法[6]においても，同様の学習用データのもと予測を行い，再現率を算出した．

表2は，予測を行った20タイトルについて，購買の可能性上位1%，2%，5%，10%および20%に対し，再現率の平均を示したものである．これを見ると，あきらかに提案手法が相関係数法を上回る高い性能となっており，本手法の有効性を確認することができる．

3.2 売上パターン予測

CDの売上は，一般に発売直後に売上が集中し，その後急速に売上が衰退していくという傾向にある．しかし，売上の減衰の仕方にはCDによって大きなバラツキが存在している．図4には，発売当初の売上はほぼ等しいものの，その後日数が経過するにしたがって比較的穏やかに売上が減少していくものと，反対に急激に減少するものの2種類の典型的な売上パターンを示した．CDが発売直後に，どちらのタイプのCDになるのかを予測することは，販売店の在庫管理上も重要な問題と考えられる．

そこで，この節ではCDの発売後1週間までの購買履歴データを用いて，その後のCDの売上パターンの予測を試みる．予測を行うにあたり，売上推移のパターンの種類として，

ロングヒット型

- 発売後1週間の売上枚数50枚以上
- 伸び率150%以上

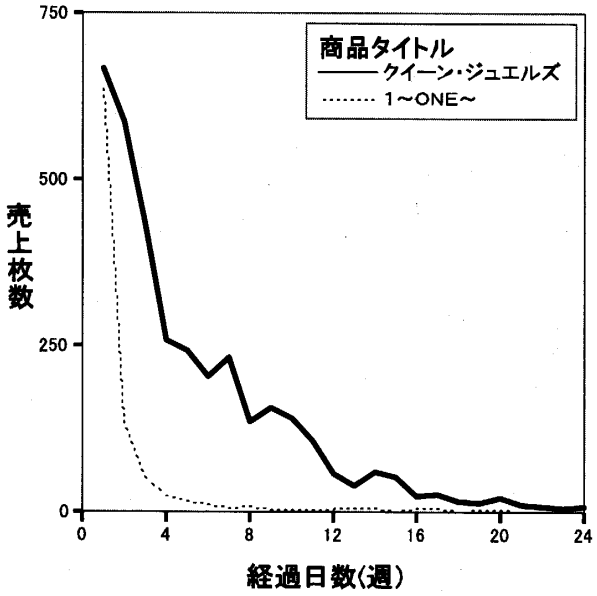

図4　CDの売上推移の様子

急降下型

- 発売後1週間の売上枚数50枚以上
- 伸び率150%未満

その他の通常型

- 発売後1週間の売上枚数50枚未満

の3種類を考える．ここで伸び率とは，

$$\frac{\text{総売上枚数}-\text{発売後1週間の売上枚数}}{\text{発売後1週間の売上枚数}}$$

で表される指標とする．この節では，比較的売上枚数の伸びが期待できるロングヒット型と急降下型に注目し，発売からちょうど1週間後の時点で累積の売上枚数が50枚以上の売れ筋のCDが，その後ロングヒットCDになるか否かを予測することを考える．

ここでの予測はCDに対するものなので，CDのみをノードとするグラフで考えれば十分である．そこで，CD間の類似度を考え，それぞれのCD毎に類似度の大きなCDをk個選び枝を張ったk最近傍グラフを構築し，ラプラシアンカーネルを求めることとした．CDとCDの類似度は，そのCDを購入した顧客のパターンにより定めることとした．すなわち，各CD iに対して顧客数の次元のベクトル$\boldsymbol{a}_i$を考え，各要素はそのCDの購買の有無を0-1で表現したものとする．このようにベクトルを定めれば，例えば，CD iとjに関してベクトルの内積$\boldsymbol{a}_i^T\boldsymbol{a}_j$を求めれば，共通して購入した顧客数が計算される．本研究では，総購入者数の違いを考慮して，ベクトル$\boldsymbol{a}_i$，$\boldsymbol{a}_j$のなす角の余弦，すなわち

$$s_{ij}=\frac{\boldsymbol{a}_i^T\boldsymbol{a}_j}{\|\boldsymbol{a}_i\|\|\boldsymbol{a}_j\|}$$

をCD iとjの類似性の指標とする．もちろん，論文の冒頭で指摘したように，ほとんどのCDの間で$s_{ij}=0$となってしまうため類似性の差異が表現できず，したがって通常の分析手法であればこのような類似性の指標は有効に機能しないものと思われる．しかし，局所的な類似性に基づいた手法では，類似性の高い部分の関連性に基づきグラフを構築することになり，このような問題は生じにくいと考えられる．

本研究では，このグラフより導出されたラプラシアンカーネルを用いて，Zhou等[10]による半教師付学習による判別予測法により，ロングヒット型CDの予測を行った．Zhou等の方法では，まず，構築したグラフのノード，すなわちCDの中で，ロングヒット型CDと急降下型CDに対して次のように行列Yを定めラベルを与える．Yは2列の行列で各行がグラフのノードに対応している．Yの1列目の列ベクトルは，ロングヒット型CDに対応する要素を1，それ以外を0としたもの，また，Yの2列目の列ベクトルは，急降下型CDに対応した要素を1，それ以外を0としたものとする．その上で，前述したラプラシアンカーネル行列$\bar{G}$を用いて，

$$F=\bar{G}Y=(I-aD^{-\frac{1}{2}}BD^{-\frac{1}{2}})^{-1}Y$$

とやはり2列の行列Fを算出する．Fは，ラベルの判定を行う行列で，f_{j1}，f_{j2}をそれぞれFのj行1列目およびj行2列目の要素とすれば，ノードjのラベルは$f_{j1}>f_{j2}$であればロングヒット型CD，そうでなければ急降下型CDと予測される．なお，手法の詳細は[5]などを参照されたい．

手法の有効性を確認するため，次のような検証を行った．まず，2年間の総売上枚数が20枚以上のCD約7千作品を選び，これらのCDを2種類以上購入した顧客16万人分を用いた．いずれのCDの予測も，発売後1週間までの全購買履歴データを基にグラフを作成し，将来の売上パターンを予測した．なお，本研究では近傍数を$k=7$とし，枝の重みは全て$b_{ij}=1$とした．

表3は，2004年の1月1日から2月28日までに発売されたCDのなかで，発売後1週間の売上枚数が50枚以上であった41枚のCDについて予測を行った結果を示したものである．これを見ると，特に判別率や精度に関しては80%以上となっており，高い精度で売上パターンを判別できていることが分かる．

表3　41商品の判別予測の結果

判別率	80.5％
精度	80.0％
再現率	70.6％

4. おわりに

本研究は，CDの購買履歴データに対する分析に，グラフ構造を用いた局所的な類似性に基づいた手法を適用し，CDの特徴抽出を行った．本研究で扱ったデータのように実際に購買されている商品が上位のもののみに集中している場合，顧客の購買行動からCD間の関連性の特徴を分析することは困難である．このようなデータに対しても，グラフ構造を用いた局所的な類似性による方法であれば，CD同士の関連性の特徴を抽出することが可能であることを確認できた．

また，構築したグラフによって導出されたラプラシアンカーネルを用いて，CDの推薦および売上推移パターンの予測を試みた．CDの推薦では，既存手法と比べて高い性能を確認することができた．また，売上推移パターンの予測においても，高い精度で判別を行う予測モデルを構築することが可能と考えられる．

さらに，グラフラプラシアンは一般には非常にスパースな行列である．したがって，たとえ数十万を超えるような大規模な行列となっても，固有値計算など適切な数値解法を用いれば通常のPCでも容易に実行が可能であり，十分にスケーラビリティーのある手法である．今後は，これらの利用法の実際のCD販売店での実施などといった，実証的な検証を行いたい．

参考文献

[1] M. Belkin and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Computation*, Vol. 15, pp. 1373-1396 (2003).

[2] I. Borg and P. Groenen, *Modern Multidimensional Scaling*, Springer series in statistics, Springer-Verlag, New York (1997).

[3] F. R. Chung, *Spectral Graph Theory*, American Mathematical Society (1997).

[4] J. Ham, D. Lee, S. Mika and B. Scholkopf, A kernel view of the dimensionality reduction of manifolds, In *ICML*-04, pp. 369-376 (2004).

[5] T.-M. Huang, V. Kecman and I. Kopriva, *Kernel Based Algorithms for Mining Huge Data Sets*, Vol. 17 of *Studies in Computational Intelligence*, Springer Verlag (2006).

[6] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon and J. Riedl, GroupLens: Applying collaborative filtering to Usenet news, *Communications of the ACM*, Vol. 40, No. 3, pp. 77-87 (1997).

[7] S. T. Roweis and L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science*, Vol. 290, pp. 2323-2326 (2000).

[8] A. J. Smola and I. R. Kondor, Kernels and regularization on graphs. In B. Schölkopf and M. Warmuth, editors, *Proceedings of the Annual Conference on Computational Learning Theory*, Lecture Notes in Computer Science, Springer (2003).

[9] J. B. Tenenbaum, V. de Silva and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science*, Vol. 290, pp. 2319-2323 (2000).

[10] D. Zhou, O. Bousquet, T. N. Lal, J. Weston and B. Schölkopf, Learning with local and global consistency, *Advances in Neural Information Processing Systems*, Vol. 16, pp. 321-328 (2004).