論　文

閲覧・購買行動に着目した結婚情報サイトにおける商品間の勝敗関係の分析手法

飯塚　修平[†a)]　濵野　将司[†]　川上　和也[†]　萩原　静厳[††]
川上　登福[†††]　浜田　貴之[††††]　松尾　豊[†]

A Product Network Based on Browsing and Purchase Behavior in E-Commerce

Shuhei IITSUKA[†a)], Masashi HAMANO[†], Kazuya KAWAKAMI[†], Seigen HAGIWARA[††], Takayoshi KAWAKAMI[†††], Takayuki HAMADA[††††], and Yutaka MATSUO[†]

あらまし　データマイニングはECサイトのマーケティングにおいて有用な情報を引き出すために広く活用されている．ユーザの商品購入パターンを発見するバスケット分析や，商品に対するユーザの嗜好を予測する協調フィルタリングなどの手法は，効果的な販売戦略の立案や商品推薦に有効である．しかし，これらの手法は商品間の類似関係や競合関係は分析できるものの，商品間の購買訴求力の優劣については分析することが難しい．そこで本研究ではECサイトのログデータから観測されるユーザの閲覧行動と購買行動に着目し，購買訴求力の優劣を表した勝敗関係を分析する手法を提案する．今回は国内の結婚情報サイトを対象にして，ログデータから商品間の勝敗関係を抽出した．評価実験では，ウェブサイトのユーザに対するアンケート結果から抽出した勝敗関係と，ウェブサイトのログデータから抽出した勝敗関係を比較した．その結果，提案手法を用いることで，アンケートデータから得られる商品間の優劣関係をログデータから推定できることがわかった．提案手法はサーバログからユーザの閲覧・購買行動及び商品に関する言語情報が取得できれば適用可能な汎用的な分析手法である．

キーワード　ECサイト，データマイニング，商品ネットワーク

1. まえがき

インターネットの普及に伴いECサイトの市場規模は拡大しており，我々の消費行動において大きな役割を果たしている．その中でデータマイニングはECサイトのマーケティングに有用な情報を引き出すための手法として広く用いられている．ユーザの商品購入パターンを発見するバスケット分析や，商品に対するユーザの嗜好を予測する協調フィルタリングなどの手法は，商品の販売戦略立案や推薦エンジンの構築に活用されている [1], [2].

しかし，これらの分析手法は同時に購入された，若しくは同じユーザに購入されたといった商品の共起関係を用いることが多い．そのため，商品間の類似関係や競合関係については分析することができるものの，各商品の購買訴求力の優劣については分析することが難しい．競合が多い市場では，競合との関係性及び優位性を認識して販売戦略を打ち出すことが重要であるため，従来の手法で得られる情報だけでは十分とは言いがたい．

そこで本研究では，ECサイトに蓄積されたログデータから商品間の競合関係及び購買訴求力の優劣関係（**勝敗関係**と呼ぶ）を分析する方法を提案する．勝敗関係は，ログデータから観測されるユーザの閲覧行動と購買行動の二つを組み合わせることで算出される．更に本研究では，勝敗関係が成立する要因すなわち勝

† 東京大学大学院工学系研究科，東京都
School of Engineering, The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–0033 Japan
†† （株）リクルートマーケティングパートナーズ，東京都
Recruit Marketing Partners Co., Ltd., 1–9–2 Marunouchi, Chiyoda-ku, Tokyo, 100–6640 Japan
††† （株）経営共創基盤，東京都
Industrial Growth Platform, Inc., 1–9–2 Marunouchi, Chiyoda-ku, Tokyo, 100–6617 Japan
†††† （株）IGPIビジネスアナリティクス&インテリジェンス，東京都
IGPI Business Analytics & Intelligence, Inc., 1–9–2 Marunouchi, Chiyoda-ku, Tokyo, 100–6617 Japan
a) E-mail: iitsuka@weblab.t.u-tokyo.ac.jp
DOI:10.14923/transinfj.2015JDP7050

者の優位性を表す**勝因キーワード**を分析する方法も提案する．勝因キーワードは，商品に関連付けられた商品情報及びレビュー文などの言語情報から算出される．

EC サイト事業者は商品間の勝敗関係及び勝因キーワードに着目することで各商品が押し出すべきアピールポイントを把握することができるようになる．また，製品開発事業者も競合市場にある天敵商品に対する自社商品の市場での強み・弱みを認識することで，今後の商品開発に役立てることができる．

今回は結婚情報サイト「ゼクシィ(注1)」で収集されたログデータを用いて商品間の関係分析を行った．まず，ログデータから観測されるユーザの閲覧行動と購買行動から，商品間の競合関係を表した競合ネットワーク及び勝敗関係を表した勝敗ネットワークを構築した．更に各商品に寄せられたレビュー文を用いて，勝因キーワードの分析を行った．

評価実験では，実際にウェブサイトを利用したユーザへのアンケート結果に提案手法を適用して算出した商品間関係と，ログデータに提案手法を適用して算出した商品間関係を比較した．その結果，この二つには有意な相関が見られた．したがって，本研究が提案する商品間関係を推定する上では，ログデータがアンケート結果の良い代替となりうることを示すことができた．また，商品のレビュー文から勝因を表すキーワードを抽出することができる可能性を示した．提案手法はユーザの閲覧・購買行動が観測できるログデータ及び商品に付加された言語情報さえあれば，EC サイト全般に導入することができる汎用的な分析手法である．

本研究の貢献は以下のとおりである．

- 閲覧及び購買ログデータから競合関係，勝敗関係及び勝因キーワードを分析する手法を提案した．
- 実際の EC サイトのログデータに提案手法を適用して分析を行った．
- 本研究が提案する商品間関係を推定する上で，ログデータがアンケートデータの代替となることを評価した．

本論文の構成は以下のとおりである．**2.** で提案手法を説明し，**3.** では提案手法を実際の EC サイトのログデータに適用した分析結果を説明する．**4.** ではアンケート結果に提案手法を用いることで得られる商品間関係を，ウェブ上のログデータから推定できることを評価する．**5.** で考察を行い，**6.** を本論文のまとめとする．

(注1)：ゼクシィ http://zexy.net

2. 提案手法

EC サイトに対してよく用いられるデータマイニング手法に，バスケット分析 [1] や協調フィルタリング [2] がある．バスケット分析では，同じユーザに閲覧若しくは購買された商品の共起関係に着目することで，商品間の関連性を導出する [3]．しかし，着目するのはあくまで選ばれた商品同士の関係であるため，選ばれなかった商品との優劣関係については分析することができない．一方，協調フィルタリングは，ある商品に対するユーザの評価を予測する．ユーザごとに商品に与えられる評価を予測するということは，そのユーザに限った商品間の優劣関係を分析する手法であると考えることもできる．しかし，こちらも同様に選ばれた商品のデータのみを用いて分析を行うため，検討されたが選ばれなかった商品の特定は困難である．したがって，いずれの手法でも選ばれた商品と選ばれなかった商品の間の購買訴求力の優劣を把握することは難しい．また，実店舗では POS データから購入された商品を把握することはできても，顧客が目にした商品や手にした商品を把握することは難しい．そのため従来の研究では，顧客に検討されたが選ばれなかった商品が着目されることは少なかった．

しかし，EC サイトは実店舗と違い，ログデータからユーザの購買行動と閲覧行動の両方を観測することができる．この 2 種類のデータを組み合わせて用いることによって，選ばれた商品と検討されたが選ばれなかった商品を導出し，その間の優劣関係を定義することができるはずである．そこで本章では，EC サイトのログデータから商品間の購買訴求力の優劣関係及びその原因を分析する手法を提案する．

2.1 競合関係と勝敗関係の定式化

ある商品を購入する際，ユーザは複数の商品を閲覧して比較検討して購入する商品を決定する．あるニーズをもったユーザによって同時に比較検討されるということは，それらの商品はユーザに提供する価値が類似しているということであり，競合となりうる商品群であると考えることができる．そこで，ここではユーザによって同時に閲覧されて比較検討される商品間の関係のことを**競合関係**と呼ぶことにする．

ユーザによる比較検討の結果，全ての商品は

- 検討されなかった商品（非閲覧，非購買）
- 検討されたが購買に至らなかった商品（閲覧，非購買）

- 検討されて購買に至った商品（閲覧，購買）

の 3 種類に区別される．検討されたが購買に至らなかった商品は，比較検討された上でユーザから「買わない」という決定を下されたので，検討された上で購入された商品に比べると購買訴求力が低いと考えることができる．検討されたが購買に至らなかった商品を**敗者商品**，検討されて購買に至った商品を**勝者商品**とすることで，商品間の購買訴求力の優劣関係を定義することができる．本研究ではこの勝者商品と敗者商品の間に結ばれる優劣関係のことを**勝敗関係**と呼ぶことにする．勝者商品同士及び敗者商品同士では勝敗関係は定義されないものとする．

ある EC サイトが取り扱う商品集合を**全商品集合** M，あるユーザ $u \in U$ によって閲覧された商品集合を**閲覧商品集合** $V_u \subset M$，購買された商品集合を**購買商品集合** $P_u \subset V_u$ とする．このとき，**勝者商品集合** $W_u = P_u$，**敗者商品集合** $L_u = V_u \setminus P_u$ が定義される．競合関係は閲覧商品集合 V_u に含まれる任意の商品間で結ばれ，勝敗関係は任意の敗者商品から任意の勝者商品に向かって結ばれる．したがって，競合関係の集合は $C_u = \{\{i,j\}|\forall i,j \in V_u, i \neq j\}$，勝敗関係の集合は $B_u = \{(l,w)|\forall l \in L_u, \forall w \in W_u\}$ と表される．

例えば図 1 のように，ある EC サイトが商品 $M = \{m_A, \ldots, m_F\}$ を取り扱っており，あるユーザ u が商品 $V_u = \{m_A, m_B, m_C, m_D\}$ を閲覧した上で商品 $P_u = \{m_A, m_B\}$ を購買した状況を考えてみる．このとき勝者商品集合は $W_u = \{m_A, m_B\}$，敗者商品集合は $L_u = \{m_C, m_D\}$ であり，六つの競合関係と $B_u = \{(m_C, m_A), (m_C, m_B), (m_D, m_A), (m_D, m_B)\}$ の四つの勝敗関係が成立する．

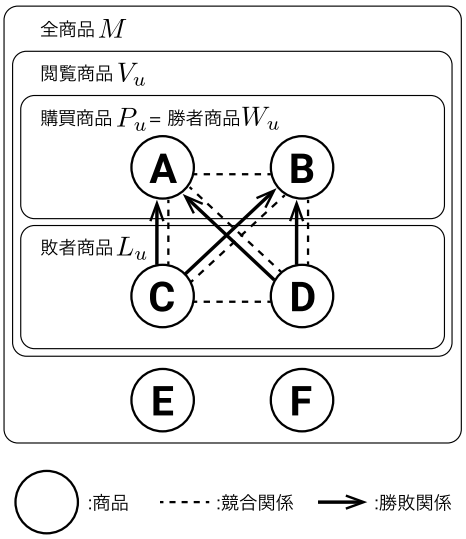

図 1　ユーザ u における商品間の勝敗関係の定義
Fig. 1　The definition of the win-lose relationship between items about user u.

2.2　競合・勝敗ネットワークの構築

商品間の競合・勝敗関係をもとに商品ネットワークを構築することで，商品間の関係の全体像を可視化することができる．また，ネットワーク科学の知見を応用することで，同じユーザに比較検討されることが多い商品のクラスタや三つ巴のように特徴的な商品間の優劣関係の発見につなげることができる．ここでは，商品間の競合関係を表したネットワークのことを**競合ネットワーク**，勝敗関係を表したネットワークのことを**勝敗ネットワーク**と呼ぶことにして，その構築方法を説明する．

競合ネットワークは各商品 $m \in M$ をノード，競合関係 $C = \bigcup_{u \in U} C_u$ をエッジとする無向グラフ $G = (M, C)$ で表すことができる．ただし，ノード m_i から m_j に向かうエッジの重みはノード間に競合関係が成り立つ回数，すなわち $w_{ij}^G = |\{u \in U|\{m_i, m_j\} \in C_u\}|$ とする．一方，勝敗ネットワークは各商品 $m \in M$ をノード，勝敗関係 $B = \bigcup_{u \in U} B_u$ をエッジとする有向グラフ $H = (M, B)$ で表すことができる．ただし，ノード m_i から m_j に向かうエッジの重みはノード間に勝敗関係が成り立つ回数，すなわち $w_{ij}^H = |\{u \in U|(m_i, m_j) \in B_u\}|$ とする．例えば 3 人のユーザ u_1, u_2, u_3 が商品 m_A, m_B, m_C について表 1 に示すような敗者商品集合 L_u 及び勝者商品集合 W_u をもつ場合，商品間の競合ネットワークは図 2 のように，勝敗ネットワークは図 3 のように表

表 1　勝敗関係の例
Table 1　Win-lose relationship example.

ユーザ u	敗者商品集合 L_u	勝者商品集合 W_u
u_1	m_B	m_C
u_2	m_A, m_B	m_C
u_3	m_C	m_B

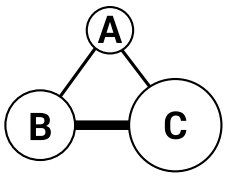

図 2　競合ネットワークの例
Fig. 2　Competitive network example.

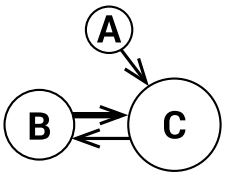

図 3　勝敗ネットワークの例
Fig. 3　Win-lose network example.

すことができる．ただし，エッジの太さはエッジの重みを，ノードの大きさは商品 m を購入したユーザ数 $p_m = |\{u \in U | m \in P_u\}|$ を表している．

2.3　勝因キーワード分析

効果的な販売戦略を打ち出すためには，単に商品間の勝敗関係を把握するだけではなく，その背後にある勝因の認識が有効である．各商品の勝因及び敗因を認識することで，商品の弱みを克服する戦略や弱みを把握した上で他の強みを伸ばす戦略などを打ち出すことができるようになる．

ここでは，商品に付加された商品情報やレビュー文などの文章で表される言語情報を活用することで，勝因を表すキーワード（**勝因キーワード**とよぶ）を分析する方法を提案する．各商品を単語の集合として表現して，勝敗関係がある商品間の差分を見ることで，勝者商品及び敗者商品の特徴を推定することができる．勝因キーワード分析は，**商品モデリング**と**勝因抽出**の二つのステップからなる．

商品モデリングとは，商品に付加された言語情報からキーワードを抽出することで，商品の特徴を単語の集合によって表すことである．ここでは，商品の言語情報を形態素解析して各商品を単語の集合として表し，tf-idf 法を用いてキーワードを抽出する方法を説明する．商品 $m \in M$ の言語情報を形態素解析して得られる単語集合を T_m とする．ここで，ある単語 k が商品 m の言語情報に出現する頻度を $tf_{k,m}$ とする．一方，単語 k を単語集号に含む商品の集合を $M_k = \{m \in M | k \in T_m\}$ として，単語 k の逆文章頻度を $idf_k = \log(|M|/|M_k|)$ と表すことにする．このとき，単語 k の商品 m における tf-idf 値は $tfidf_{k,m} = tf_{k,m} \cdot idf_k$ と表すことができる．ここでは，各商品 m について tf-idf 値が高い上位のキーワードをその商品の特徴を表すキーワード集合 K_m とする．

勝因抽出では，各商品 m のキーワード集合 K_m から勝因キーワードを抽出する．商品 m_i が商品 m_j に勝利したとき，その勝敗関係を $m_i \succ m_j$ と表すことにすると，商品 m_i が商品 m_j より好ましいと判断したユーザの集合は $U_{m_i \succ m_j} = \{u \in U | m_i \in W_u, m_j \in L_u\}$ と表すことができる．ユーザ $u \in U_{m_i \succ m_j}$ が閲覧した商品 $v \in V_u$ のキーワード集合 K_v は，商品 m_j より商品 m_i を好むユーザの嗜好が反映された単語が集まったものだと考えることができる．したがって，商品 m_j に対する商品 m_i の勝因キーワードは $K_{m_i \succ m_j} = \bigcup_{u \in U_{m_i \succ m_j}} \bigcup_{v \in V_u} K_v$ と算出することができる．

3.　EC サイトでの分析結果

本章では提案手法を結婚情報サイト「ゼクシィ（**EC サイト A** と呼ぶ）」で収集されたログデータに適用した例を示す．提案手法はユーザが閲覧した商品と購買した商品の差をもとに勝敗関係を算出するため，ユーザが多くの商品を閲覧して比較検討する EC サイトと相性が良い手法である．比較検討される商品が多くなるほど多くの勝敗関係を定義することができるようになるため，分析がより精緻なものになると考えられる．そこで本研究では単価が大きく，ユーザの事前知識が少ないために慎重な比較検討を要すると考えられる結婚式場を商品として取り扱った EC サイト A を対象サイトとして選定した．EC サイト A は情報提供サイトであり厳密には EC サイトではないが，ユーザはこのサイトを閲覧して商品の情報を収集して見学予約や問合せなどの購買につながる行動を起こしているため，本研究では EC サイト A を広義の EC サイトとみなして分析することにする．

EC サイト A は国内の結婚情報サイトであり，サイト上で結婚式場や結婚指輪の詳細ページを閲覧したり，結婚式場の見学予約や結婚指輪の資料請求をしたりすることができる．本分析では，EC サイト A で取り扱う結婚式場を商品，結婚式場の詳細ページの閲覧を閲覧行動，結婚式場の見学予約を購買行動とみなして分析を行う．分析に用いるデータセットは，EC サイト A において 2012 年 1 月 1 日から 2012 年 10 月 31 日の 10ヵ月間に収集されたログデータである．ログデータには匿名化されたセッション ID と閲覧された式場，そして見学予約の有無が含まれている．ここではセッション ID をユーザに割り振られた識別子として扱うことにする．この期間中に数百万件のセッションによる式場閲覧，数万件の見学予約が確認された．以下では，これらのセッションのうち期間中に式場閲覧と見学予約の両方が観測されたものを用いて勝敗関係の分析を行う．

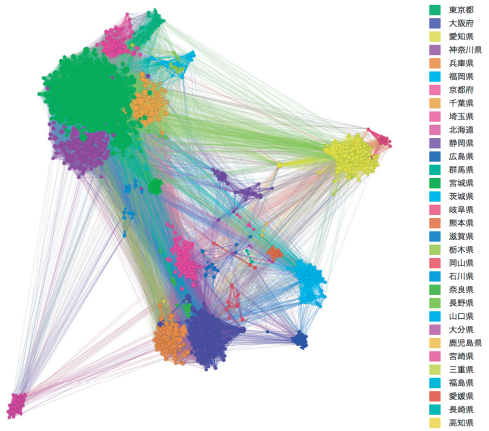

図 4　全国の式場の競合ネットワーク（都道府県別に色分け）

Fig. 4　The competitive network of ceremonial halls around Japan (colored by prefectures).

図 4 に全国の式場の競合ネットワークを示す．ノードは全国の式場 M，エッジは式場間の競合関係 C，ノードの大きさは式場 m の累計購買回数 p_m，エッジの太さはエッジの重み w_{ij}^G を表しており，都道府県別にノードの色を分けている．ただし累計閲覧回数 $v_m = |\{u \in U | m \in V_u\}|$ があるしきい値に満たない式場は描写を省略しており，その結果約 1,000 個のノードと約 75,000 本のエッジが描写されている．ネットワークは力学モデルによってレイアウトされており，エッジをばねとみなした引力と，ノードを電荷をもつ粒子とみなした斥力が平衡した状態が描写されている [4]．そのため，太いエッジが張られているほどノード間の距離は近くなり，エッジが張られていなければ斥力によって距離が遠くなる．ノードの色分けがネットワーク上のクラスタに対応していることから，各式場は他の都道府県に存在する式場を競合として考える必要性が低いことがわかる．

図 5 と図 6 に東京都内の式場の競合ネットワークを示す．図 5 はエリア別に，図 6 は業態別に式場ノードの色を分けている．ただし累計閲覧回数 $v_m = |\{u \in U | m \in V_u\}|$ があるしきい値に満たない式場は描写を省略しており，その結果約 300 個のノードと約 25,000 本のエッジが描写されている．図 5 ではノードの色分けがネットワーク上のクラスタに対応していないが，図 6 ではノードの色分けがネットワーク上のクラスタに対応していることがわかる．このことから，東京都内の式場は地理的な条件よりも業態に着目して競合式場を想定する必要性が高いことがわかる．

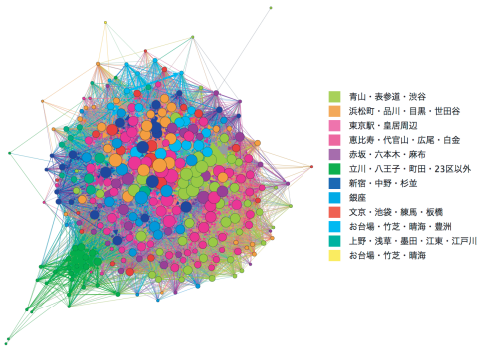

図 5　東京都内の式場の競合ネットワーク（エリア別に色分け）

Fig. 5　The competitive network of ceremonial halls in Tokyo (colored by areas).

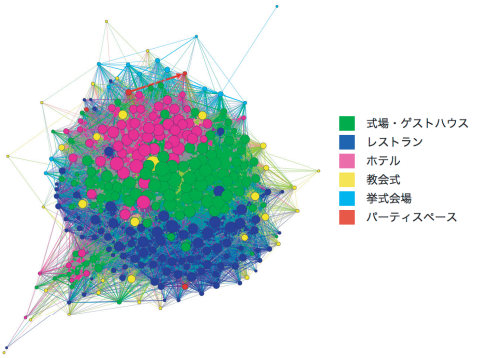

図 6　東京都内の式場の競合ネットワーク（業態別に色分け）

Fig. 6　The competitive network of ceremonial halls in Tokyo (colored by business categories).

表 2　選出した 10 式場のエリアと業態

Table 2　Areas and business categories of the selected 10 halls.

式場	エリア	業態
A	浜松町・品川・目黒・世田谷	式場・ゲストハウス
B	文京・池袋・練馬・板橋	ホテル
C	恵比寿・代官山・広尾・白金	式場・ゲストハウス
D	青山・表参道・渋谷	式場・ゲストハウス
E	文京・池袋・練馬・板橋	式場・ゲストハウス
F	東京駅・皇居周辺	ホテル
G	青山・表参道・渋谷	式場・ゲストハウス
H	赤坂・六本木・麻布	式場・ゲストハウス
I	お台場・竹芝・晴海・豊洲	ホテル
J	浜松町・品川・目黒・世田谷	ホテル

次に，表 2 に示す東京都内の 10 式場（式場 A～J）を取り上げて勝敗ネットワークを構築したものを図 7 に示す．式場 A, C, E, H 間を結ぶエッジは特に太く，

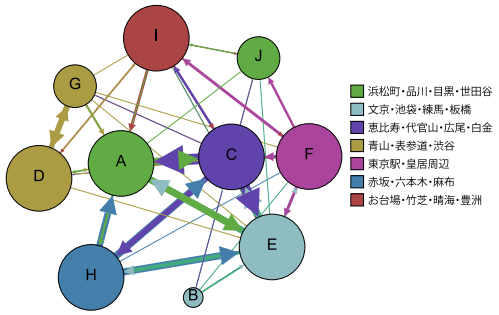

図 7　10 式場の勝敗ネットワーク（エリア別に色分け）
Fig. 7　The win-lose network of the 10 halls (colored by areas).

表 3　式場 A の式場 C, H に対する勝因キーワード
Table 3　The keywords that indicate the advantage of hall A against hall C and hall H.

式場 C に対する勝因	式場 H に対する勝因
コストパフォーマンス	コストパフォーマンス
招待	丁寧
参加	招待
試食	試食
チャペル	残念
普通	ゲスト
感動	人数
立地	相談
丁寧	プランナー
残念	立地

これらは強い競合関係にあることがわかる．ここで式場 H に着目すると式場 A については入次数よりも出次数が大きくなっており，負け越していることがわかる．したがって式場 H は式場 A, C, E を主な競合として注意しながら，天敵となる式場 A とは特に優位性を意識した販売戦略を立てることが必要だとわかる．

最後に，取り上げた 10 式場でも特にノードが大きく，強い競合関係にある式場 A, C, H に着目して勝因キーワード分析を行った．今回は式場に付加されたレビュー文として，国内の結婚式場口コミサイトに寄せられた東京都内の式場に対する約 25,000 件のレビュー文を用いた．また，分析対象の式場の 75%以上の式場のレビュー文に出現する単語は極端に文章頻度が大きい単語と判断して除外した．

表 3 に式場 A の式場 C, H に対する勝因キーワードを出現頻度が多い順に 10 個並べたものを示す．「コストパフォーマンス」は両方の式場に対して最上位に挙げられており，式場 A の強みを表したキーワードであると考えられる．式場 C に対しては「チャペル」「感動」といったキーワードが挙がっていることから，チャペルを使った感動的な演出が相対的な強みとなっていると考えられる．また，式場 H に対しては「相談」「プランナー」といったキーワードが上がっており，結婚式のプラニングの丁寧さで相対的に優っていると推察される．

表 4　アンケートデータ D と EC サイト A におけるログデータ $\hat{D}$ の対応関係
Table 4　The correspondence between the questionnaire result data D and log data $\hat{D}$.

	アンケートデータ D	ログデータ $\hat{D}$
ユーザ集合	回答者 (N=173)	訪問ユーザ (N=202)
閲覧行動	式場見学	式場詳細ページの閲覧
購買行動	挙式する式場の決定	見学予約
言語情報	挙式した式場の選択理由	レビュー文

4. 評価実験

本章では，ログデータに提案手法を適用して算出した商品間関係と，EC サイト A を利用したユーザに対して行ったアンケートデータに提案手法を適用して算出した商品間関係を比較し，ログデータがアンケートデータの良い代替となることを評価する．このアンケートは EC サイト A を用いて見学予約を行い，挙式する式場を決定したユーザを対象としたもので，比較検討した式場，実際に挙式した式場及びその選択理由が記載されている．ユーザが明示的に回答したものであるため，このアンケートデータには式場選択に関するユーザの嗜好が反映されていると考えられる．一般に，アンケートデータを得るには回答者の募集や調査票の準備など手間がかかることが多い．一方，ログデータはユーザに特別な負担をかけることなく容易に収集することができるデータである．そのため，商品間関係を知るうえでログデータでアンケートデータを代替することができれば，その利用価値は高いと考えられる．そこで本実験では，アンケートデータに提案手法を適用することで得られる商品間関係を**真の商品間関係**と仮定し，ログデータに提案手法を適用することで真の商品間関係を推定できることを評価する．

アンケートデータとログデータの対応関係を表 4 に示す．比較検討した式場が閲覧行動，実際に挙式した式場が購買行動，そして選択理由が式場に付与されたレビュー文と対応しているため，アンケートデータにも提案手法を適用して商品間関係を抽出することが可能である．分析対象は **3.** で選出した東京都内の 10 式場とし，ログデータも **3.** と同じデータセットを用いる．

まず，商品間の競合関係について評価を行う．対象となる 10 式場の組み合わせ 45 個について，アンケートデータ D とログデータ $\hat{D}$ からそれぞれ真の競合関係の重み w^G と推定競合関係の重み $\hat{w^G}$ を算出する．ここでは，ある商品間の競合関係は他の商品間の競合関係とは独立に決定されるものと仮定する．それぞれの重み同士の相関分析を行った結果，Pearson の相関係数は 0.685，p 値は 2.06×10^{-7} となり，有意な相関が認められた．このとき，40 個のうち 6 個の組み合わせについては真の競合関係の重みまたは推定競合関係の重みの値が 0 であったが，それ以外の組み合わせについては両方に重みが割り当てられていた．

次に，商品間の勝敗関係について評価を行う．勝敗関係は競合関係と異なり方向があるので，対象の 10 式場間には 90 個の勝敗関係が結ばれる．ここでは，ある商品間の勝敗関係は他の商品間の勝敗関係とは独立に決定されるものと仮定する．それぞれの勝敗関係について，アンケートデータ D とログデータ $\hat{D}$ から真の勝敗関係の重み w^H と推定勝敗関係の重み $\hat{w^H}$ を算出する．その後，それぞれの重み同士の相関分析を行った結果，Pearson の相関係数は 0.648, p 値は 5.02×10^{-12} となり，有意な相関が認められた．このとき，90 個のうち 20 個の勝敗関係については真の勝敗関係の重みまたは推定勝敗関係の重みの値が 0 であったが，それ以外の組み合わせについては両方に重みが割り当てられていた．

最後に，勝因キーワード分析について評価を行う．まず，アンケートデータ D に記された挙式式場の選択理由から真の勝因キーワード K を算出する方法について述べる．挙式した式場の選択理由には，見学されたが選ばれなかった敗者式場と選ばれた勝者式場間の勝敗関係の要因が記されていると考えられる．そこで，ここでは選択理由に対して形態素解析を行い，一般名詞，サ変接続名詞または形容動詞語幹名詞であり，かつ出現頻度 2 以上の単語を，アンケートを記入したユーザ u の嗜好を表したキーワード K_u とする．このキーワード K_u を商品 m_j を検討したが商品 m_i で挙式を挙げたユーザ $u \in U_{m_i \succ m_j}$ について足しあわせたものを，商品 m_j に対する商品 m_i の真の勝因キーワード $K_{m_i \succ m_j} = \bigcup_{u \in U_{m_i \succ m_j}} K_u$ と定義する．

本実験では，提案手法によってログデータ $\hat{D}$ から算出された勝因キーワード $\hat{K}$ と真の勝因キーワード K の一致数 $|K \cap \hat{K}|$ によって推定の有効性を評価する．比較手法として，レビュー文から勝因抽出は行わず商品モデリングによって商品 m のキーワード抽出のみを行ったもの $\hat{K'}_m$ を設定する．このキーワードは，他の全ての商品に対する商品 m の勝因キーワードとして扱うことにする．提案手法と比較手法いずれも td-idf 値上位 50 個の単語を勝因キーワードとして採用することにする．

表 5　勝因キーワード分析における提案手法と比較手法の性能比較
Table 5　The performance comparison between the baseline method and the proposed method on the cause-of-win keyword analysis.

ペア	A vs C		A vs H		C vs H		合計
勝者式場 m_i	A	C	A	H	C	H	
敗者式場 m_j	C	A	H	A	H	C	
真の KW 数 $\lvert K_{m_i \succ m_j} \rvert$	22	51	10	14	27	14	138
提案一致数 $\lvert K_{m_i \succ m_j} \cap \hat{K}_{m_i \succ m_j} \rvert$	8	18	3	4	9	4	46
比較一致数 $\lvert K_{m_i \succ m_j} \cap \hat{K'}_{m_i} \rvert$	6	17	4	6	10	7	50

真の勝因キーワード K と提案手法及び比較手法によって推定された勝因キーワード $\hat{K}, \hat{K'}$ の一致数を比較した結果を表 5 に示す．全パターンについての真の勝因キーワードとの一致数の合計は，提案手法が 46，比較手法が 50 となり，提案手法は比較手法を下回った．式場 A と式場 C のペアにおける勝因キーワードの一致数は提案手法が比較手法を上回っていたが，それ以外のペアについてはいずれも比較手法が提案手法を上回る結果となった．

5. 考　　察

5.1　提案手法の有用性と応用可能性

ログデータから得られた商品間関係とアンケートデータから得られた商品間関係を比較した結果，競合関係・勝敗関係ともに有意な相関が見られた．したがって，本研究で提案する商品間の優劣関係を用いる限りは，ログデータがアンケートデータの良い代替になると考えられる．つまり，実際にユーザに働きかけて直接意見を得ることが難しい場合でも，ウェブサイト上の行動から暗黙的に取得されたログデータに提案手法を適用することで，ユーザが考える商品間の優劣関係を推定することができる．

勝因キーワード分析の適用例では，商品をキーワード集合で表し，勝敗関係による差分を取ることで勝因キーワードを推定することができる可能性を示した．しかし評価実験においては，提案手法は比較手法に比

べて真の勝因キーワードとの一致数が低く，真の勝因キーワードを推定するには有効な手法であるとは言えなかった．

提案手法が勝因キーワードを推定できなかった原因として，勝因抽出の方法に問題があったと考えられる．勝因分析は，ある商品間の勝敗 $m_i \succ m_j$ に寄与したユーザが閲覧した商品 $v \in V_u$ に含まれるキーワード K_v の和として定義される．これはある商品間の勝敗に寄与するユーザには一貫した嗜好があり，その和をとることによってその嗜好を表すキーワードが浮かび上がってくることを仮定している．しかし，実際には足しあわせていく過程でキーワードが増えてしまい，かえって嗜好を表すキーワードが不明確になってしまった可能性がある．閲覧行動ではなく，より強くユーザの嗜好を表していると考えられる購買行動を用いて勝因抽出を行う工夫が考えられる．または，キーワード抽出を行わずに単語に直接勝因抽出を行うことで，出現頻度は低いが重要な特徴を表す単語が弾かれてしまうことを防ぐ工夫も考えられる．

商品点数が極端に多い場合にはネットワークが疎になる可能性があるため，購買回数が少ない商品を省く，若しくは類似商品をマージするなどの工夫が必要になると考えられる．逆に商品点数が極端に少ない場合には，ノード間に張られるエッジの数が増えて商品間の関係性を把握することが困難になることが考えられる．この場合には，あるしきい値以下の重さのエッジを取り除く工夫が考えられる．

ユーザ数が極端に多い場合には多くの商品間関係を抽出することができるが，その裏に様々な異なる嗜好が反映されてしまう可能性がある．この場合，ユーザの属性ごとに商品間関係を分析することによって精緻な販売戦略の立案が可能になると考えられる．逆にユーザ数が極端に少ない場合にも，提案手法は閲覧行動で表されるユーザの暗黙的な嗜好データを元にしているため，購買行動をもとに商品間関係を分析する場合に比べると多くの商品間関係を抽出することができると考えられる．

結婚式場や不動産のように，単価が高く慎重な購買行動が求められる商品カテゴリーは比較検討する商品が多いため，提案手法との相性が良いと考えられる．一方，書籍やゲームなどのコンテンツ産業商品やビールや洗剤などの寡占市場の商品は比較検討が十分に行われないため，十分な量の商品間関係を抽出できない可能性がある．

対象の EC サイトでキャンペーンを行った場合には，その期間キャンペーン対象の商品が勝者になる傾向が強くなる可能性がある．しかし，キャンペーン目的で来訪したユーザは，他の商品との比較検討を行わずにキャンペーン対象の商品を購買するため，敗者商品も少なくなると考えられる．したがって，提案手法はキャンペーンや流行などの外部要因にも強い手法である．

本研究の提案手法は商品間の競合関係の定義，商品間の勝敗関係の定義，勝因キーワード分析の三つで構成されている．閲覧行動，購買行動にまつわるログデータと商品に付加された言語情報の 3 種類のデータさえあれば適用可能な手法であるため，提案手法はあらゆる EC サイトのログデータに適用できる汎用性が高い手法であると考えられる．商品に付加された言語情報として本研究ではレビュー文を用いたが，商品説明文や商品の属性値などのデータでも代用することができる．

5.2 関 連 研 究

バスケット分析は商品間の関係分析のために実店舗の POS データ及び EC サイトのログデータに対して適用されている [1]．アプリオリ・アルゴリズムの登場によって大規模データに対しても適用できるようになり，広く活用されるようになった [5]．バスケット分析を商品ネットワークの構築に活用する研究も進められており，相関ルールをもとに商品を有向エッジで接続して可視化する試みや [6]，そこから商品のコミュニティを抽出して俯瞰する試みがなされている [7]．他にも商品ネットワークにスケールフリー性が見られること [8]，着目する共起関係によってネットワークの性質が変化することが報告されている [9]．一方，EC サイトのログデータから観測されるユーザの閲覧行動と購買行動から商品間の優劣関係を定義する試みもあり，優劣関係に着目することが商品推薦のランク付けに有効であることが示されている [10]．

このように，商品の共起を元に商品ネットワークを構築する試み，EC サイトの閲覧履歴と購買履歴から商品間の優劣関係を定義する試みは行われているが，これらを組み合わせて商品間の勝敗関係に基づいたネットワークを構築し，その性質を実際のデータを用いて評価したところが本研究の新規性である．

6. む す び

本研究では，EC サイトの閲覧行動と購買行動に関するログデータ及び商品に付加された言語情報を用い

て，商品間の競合関係と勝敗関係及び勝因キーワードの分析を行う手法を提案した．今回は国内の結婚情報サイト「ゼクシィ」に提案手法を適用して分析を行った．評価実験では，サイトを利用したユーザを対象にしたアンケート結果に提案手法を適用して抽出した真の商品間関係と，ログデータから抽出した商品間関係を比較した．その結果，競合関係・勝敗関係ともに有意に相関しており，本研究が提案する商品間の優劣関係を推定する上では，ログデータがアンケートデータの良い代替になることが示された．勝因キーワード分析についても同様にアンケート結果から得られる真の勝因キーワードの推定を行った結果，提案手法は比較手法に及ばなかったが，商品間の勝敗関係の原因を表すキーワードを抽出できる可能性を示すことができた．提案手法は EC サイトで一般的に収集することができる 3 種類のデータがあれば適用することができる汎用的な分析手法である．

文　　献

[1] Y.-L. Chen, K. Tang, R.-J. Shen, and Y.-H. Hu, “Market basket analysis in a multiple store environment,” Decision Support Systems, vol.40, no.2, pp.339–354, 2005.

[2] 土方嘉徳，“嗜好抽出と情報推薦技術，” 情報処理，vol.48, no.9, pp.957–965, 2007.

[3] S. Brin, R. Motwani, and C. Silverstein, “Beyond market baskets: Generalizing association rules to correlations,” ACM SIGMOD Record, vol.26, pp.265–276, 1997.

[4] M. Jacomy, S. Heymann, T. Venturini, and M. Bastian, “Forceatlas2, a continuous graph layout algorithm for handy network visualization,” Medialab Center of Research, 2011.

[5] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” Proc. 20th Int. Conf. Very Large Data Bases, VLDB, vol.1215, pp.487–499, 1994.

[6] M.C. Hao, U. Dayal, M. Hsu, T. Sprenger, and M.H. Gross, Visualization of directed associations in e-commerce transaction data, Springer, 2001.

[7] I.F. Videla-Cavieres and S.A. Ríos, “Extending market basket analysis with graph mining techniques: A real case,” Expert Systems with Applications, vol.41, no.4, pp.1928–1936, 2014.

[8] T. Raeder and N.V. Chawla, “Modeling a store’s product space as a social network,” International Conference on Advances in Social Network Analysis and Mining, 2009, ASONAM ’09, pp.164–169, 2009.

[9] H.K. Kim, J.K. Kim, and Q.Y. Chen, “A product network analysis for extending the market basket analysis,” Expert Systems with Applications, vol.39, no.8, pp.7403–7410, 2012.

[10] 武政孝師，後藤順哉，“EC サイトにおける顧客の閲覧履歴を利用した商品ランキング生成法，” オペレーションズ・リサーチ，vol.59, no.8, pp.465–471, 2014.

（平成 27 年 5 月 10 日受付，8 月 18 日再受付，
9 月 28 日早期公開）

飯塚　修平　（学生員）

2012 年 3 月東京大学工学部システム創成学科知能社会システムコース卒業．2014 年 3 月同大大学院工学系研究科技術経営戦略学専攻修士課程修了．現在，同専攻博士課程在学中．2014 年より Google 株式会社 UX エンジニア．

濵野　将司

2012 年 3 月東京大学薬学部薬科学科卒業．2014 年 3 月同大大学院工学系研究科技術経営戦略学専攻修士課程修了．

川上　和也

2014 年 3 月東京大学卒業，2014 年 8 月よりカーネギーメロン大学在学中．専門は機械学習，自然言語処理．

萩原　静厳

2003 年東京工業大学工学部卒業．2005 年同大大学院修士課程修了．同年より株式会社リクルートにて旅行雑誌「じゃらん」広告営業，インターネット宿泊予約サイト「じゃらん net」企画及び UX デザイン業務．2009 年 4 月より結婚情報サイト「ゼクシィ」企画，UX デザイン及びビッグデータ解析業務．2014 年 10 月よりオンライン教育サービス「受験サプリ」「勉強サプリ」にてデータ解析及び UX デザイン業務に従事．現在ビッグデータエバンジェリスト．2011 年度ブライダルカンパニー年間最優秀賞（MVS）など受賞．専門は UX デザイン，データアナリティクス，Web サービス戦略企画．

川上　登福

大手商社，GE にて，営業・マーケティング，企業再生，会社立ち上げ，M&A，JV・業務提携等に従事．IGPI 参画後は，大手メディア・IT 系企業やインフラ企業等の戦略策定・新規事業開発・ハンズオン支援や，データ解析・アルゴリズム開発，データ活用戦略策定等に従事．NEDO 次世代ロボット中核技術開発/次世代人工知能技術分野採択審査委員．立命館大学法学部卒．現在，株式会社経営共創基盤パートナー取締役マネージングディレクター兼株式会社 IGPI ビジネスアナリティクス&インテリジェンス代表取締役 CEO．

浜田　貴之

大手日系・外資系証券会社にて，金融工学を用いたデリバティブ等の金融商品設計業務・財務戦略立案・M&A に従事．IGPI 参画後は，製造業・金融・ネット系企業を中心に，事業計画の策定・実行支援，市場規模・売上予測，KPI 管理・管理会計の仕組み構築，ビッグデータの経営判断への活用支援，各種アルゴリズムを活用した新規事業・新サービス開発等に従事．University of California, Berkeley B.S. Mechanical Engineering. New York University, Stern School of Management M.S. Business Analytics．現在，株式会社 IGPI ビジネスアナリティクス&インテリジェンス代表取締役 COO．

松尾　　豊

1997 年東京大学工学部卒業．2002 年同大大学院博士課程修了．博士（工学）．産業技術総合研究所，スタンフォード大学を経て，2007 年より，東京大学大学院工学系研究科技術経営戦略学専攻准教授．2012 年より人工知能学会理事・編集委員長，2014 年より倫理委員長．人工知能学会論文賞，情報処理学会長尾真記念特別賞，ドコモモバイルサイエンス賞など受賞．専門は，Web 工学，Deep Learning，人工知能．