『学習院大学　経済論集』第48巻　第４号（2012年１月）

ソーシャルメディアを対象としたマーケティング解析
―時事問題をきっかけとした想定外の消費行動抽出―

橋本　隆子*，久保山　哲二**，白田　由香利***

Abstract.

ブログや口コミ掲示版といったソーシャルメディアから製品の評判やニーズを分析する研究が盛んである。既存の研究では，特定の製品や機能に注目し，「好き」，「嫌い」，「高い」，「便利」といった典型的な評価語の発生頻度や Positive／Negative の度合いを定量化することで消費者の関心やそれに伴う消費行動の解析が行われている。しかしながら，消費者のニーズや関心は特定の製品や機能に対して直接的に示されるだけではなく，種々の時事問題を反映して間接的に示されることもあり，結果として意外な消費行動を引き起こす場合がある。時事問題をトリガーとした想定外の消費行動パターンを発見できれば，新しいマーケティングリサーチ手法となると我々は考える。そこ本論文では，口コミ掲示版の書き込みから時事問題と製品間の相関を抽出し，想定外の消費行動を発見する手法を提案する。提案手法は，まず時事問題と各種製品間の時系列相関を Dynamic Time Warping 法により算出し，時事問題との間に想定外の相関関係をもつような製品候補を抽出する。さらにその製品候補の口コミ掲示版において発生する語の共起関係をベースに消費者の書き込みをネットワーク構造化し，話題構造の推移を時系列で可視化する。時系列グラフ構造の動的な振舞いを分析することで，時事問題をきっかけとした想定外の消費行動を抽出していく。我々の手法により，時事問題に対して一見無関係に思われる製品に対する消費者の想定外の消費行動を分析することが可能となる。

1　はじめに

ブログや口コミ掲示版といったソーシャルメディアから製品の評判やニーズを分析する研究が盛んである［１－５］。これらの研究では，特定の製品や機能に注目し，「好き」，「嫌い」，「高い」，「便利」といった典型的な評価語をベースとした解析が行われている。著者らも「空気清浄機」や「ドラム式洗濯機」といった特定の家電製品を対象として，口コミ掲示版の書き込みからその特徴や評判を分析する「評判分析フレームワーク」の提案を行い［６］，適切な話題分類手法について考察してきた。一般に，製品を特定した評判解析は比較的容易である。製品

*）　千葉商科大学 商経学部　takako@cuc.ac.jp
**）学習院大学 計算機センター　kuboyama@tk.cc.gakushuin.ac.jp
***）学習院大学 経済学部　yukari.shirota@gakushuin.ac.jp

の特徴や頻出キーワードを事前に想定可能であり，辞書等も用意しやすく，対象を絞って解析できるためである。しかしながら，実際には消費者のニーズや関心は特定の製品や機能に対して直接的に示されるだけではなく，種々の時事問題を反映して間接的に示されることもある。たとえばインフルエンザの流行をきっかけとして，ウイルス除去機能を持つ空気清浄機に対して消費者が関心をもち，購買意欲が玉カルというのは容易に想定可能な消費行動である。一方，一見インフルエンザとは無関係な製品（デジタルカメラなど）に対して，インフルエンザの流行がその製品を買い控えるといった消極的な行動に影響を及ぼす場合もある。こうした時事問題をきっかけとした想定外の消費行動分析は，従来にない新しいマーケティングリサーチ手法となると我々は考える。想定外の消費行動発見するために，我々はグラフ構造を利用した可視化手法を提案してきた［7，8］。また単語の共起を時系列に評価する Dynamic Time Warping（DTW）法［9］により，時事問題と製品間の想定外の共起を発見する手法についても提案してきた［10］。本稿では，我々の従来研究を整理・統合し，ソーシャルメディアを対象とした想定外の消費行動分析のためのフレームワークを提案する。提案手法は，時事問題と各種製品名の発生頻度情報基づく時系列相関を算出し，時事問題と想定外の相関関係をもつような製品候補を抽出する。抽出された製品候補に関するソーシャルメディア上の書き込みについて，特徴語の共起関係をベースに話題構造をネットワークグラフ化し，グラフ構造の時系列推移を可視化する。時系列グラフ構造の動的な振舞いを分析することで，時事問題をきっかけとした想定外の消費行動を解析していく。

本稿は以下の構成となっている。第２章では我々が考える想定外の消費行動について述べる。第３章でソーシャルメディア解析，時系列相関，グラフ構造による話題抽出手法に関する関連研究を紹介し，我々のアプローチとの違いを説明する。第４章では時事問題をきっかけとした想定外の消費行動を抽出する，我々の提案手法について説明する。第５章では提案手法を実際のデータに適用した結果を示す。第６章で結論及び今後の展開について述べる。

２　想定外の消費行動とは

本章では，我々が考える想定外の消費行動とはどういったものかについて述べていく。まず想定内の消費行動について説明し，それに対応付けつつ想定外の消費行動について述べる。

2.1　想定内の消費行動

我々は想定内の消費行動を，時事問題と明示的な関係を持つ製品に対する消費者の行動として定義する。たとえばインフルエンザが流行した際，ウイルス除去機能をもつ空気清浄機に対して消費者の興味が高まり，口コミ掲示版での書き込みが活発になり，購買意欲が向上するというのは容易に想像できる事象である。出荷台数といったマーケティング情報と照らし合わせ，実際に空気清浄機の売り上げが伸びていることも簡単に確認できる。このようなインフルエンザの流行とウイルス除去機能をもつ空気清浄機の関係は，明確であり容易に想定可能である。こうした容易に想定可能な関係に基づく消費者の行動を，我々は想定内の消費行動と定義する。

2.2　想定外の消費行動

一方，時事問題とは一見無関係に思える製品が，実は時事問題に影響されており，時事問題の活性化をきっかけとして消費者が製品を購入に走る，あるいは買い控えるといった行動に出るという事象がある。たとえば2009年にインフルエンザが流行した際に，デジタルカメラの出荷台数が例年に比べて減少するという状況が見られた。実際に口コミ掲示版上のデジタルカメラに関するスレッドでは，インフルエンザの流行のせいで旅行をキャンセルした，子供の運動会が中心になったといった書き込みがあり，デジタルカメラに興味のある消費者がインフルエンザの流行に対して反応している様子が見て取れた。インフルエンザの流行がデジタルカメラの消費者に影響を及ぼし，カメラを買い控えるといったネガティブな消費行動が引き起こされたということが予想される。こうした時事問題とは一見無関係に見える製品が，時事問題に影響を受け，間接的に引き起こされる消費行動を我々は想定外の消費行動と定義している。想定外の消費行動はいわば隠れた状態ということができ，想定外の消費行動を検知することで，新たなマーケティングリサーチ手法を提案していくことが可能であると我々は考える。

3　関連研究

本章では，関連研究を１）ソーシャルメディアを対象とした製品の評判分析研究，２）時系列相関分析の研究，３）グラフ構造を利用した話題抽出手法，の３種類に分類し紹介を行っていく。

3.1　ソーシャルメディアを対象とした製品の評判分析

ソーシャルメディアの書き込みから製品の評判を分析する研究は数多く行われている［１－４］。Nagano らはソーシャルメディアを対象として製品の評判を可視化する口コミ解析エンジンを開発している［１］。Nagano らのシステムでは，ユーザは製品の写真を携帯端末などで撮影し，システムに送信する。システムは製品を画像解析により特定し，ソーシャルメディア上の該当製品に関する口コミ（「好き」，「嫌い」，「高い」，「便利」など）を検索する。検索結果から Positive／Negative の度合いを算出している。 Kobayashi ら［２］は，評判を（対象製品，属性，意見）の３つ組みとして表現し，データベース化を行っている。 Asano ら［３］もまた，評判を（対象製品，評価ポイント，表現）の３つ組みで表わしている。Kobayashi ら及び Asano らは，対象製品に関する属性表現や意見を示す辞書（オントロジー）を効率的に生成する手法を提案し，ソーシャルメディアから評判を抽出する手法を提案している。 Spangler ら［４］は，特定の企業のブランドイメージや評判，消費者の嗜好，消費行動を解析するためにソーシャルメディアを自動的に監視するシステムを提案している。また彼らは，監視結果に基づいて特定の評価表現の Positive／Negative の度合いを計算することで，ほぼリアルタイムにオントロジーを開発する手法についても提案を行っている。これらの関連研究は，特定の製品を対象とし，専用のオントロジーを利用してソーシャルメディアから評価表現を抽出し，その Positive／Negative 度合いをベースに評判を解析すると言うアプローチを取っている。

我々の手法は，特定の製品を対象とせず，特定のオントロジーも必要としない。我々はソーシャルメディアの書き込みを解析し，時事問題と不特定の製品間の関係にを可視化し，想定外の消費行動の発見を行っていく。可視化を行うことで，より効率的に想定外の消費行動を発見

することが可能となる。

3.2　時系列相関分析の研究

時系列で相関を分析する研究も種々行われている。Zhu ら［11］は利用者のハミングによる楽曲検索を目的として，Dynamic Time Warping（DTW）法［９］を活用した時系列相関による楽曲検索手法を提案している。 Otanto ら［12］は２つの時系列データ間の相関を 動的に解析するために Dynamic Conditional Correlation model を提案している。彼らは特に時系列経済データを対象としている。Loy ら［13］は複数のカメラによって撮影された映像を時系列で評価解析することにより，さまざまな動作を理解する手法を提案している。彼らは Cross Canonical Correlation Analysis（xCCCA）を利用し，複数のカメラ映像における時系列相関を発見する手法を定式化している。

我々の想定外の消費行動抽出手法は DTW を活用しており，その点では Zhu らのアプローチと同じであると言える。しかしながら，我々の手法はソーシャルメディアにおける時事問題（単語）と製品名，及び各種内容語の共起に注目し，その相関関係を解析することで想定外の消費行動を発見する手掛かりとしている。対象とするデータが異なっており，言語処理と組み合わせている点で従来研究とはアプローチが違うと言える。

3.3　時系列による単語共起分析の研究

グラフ構造を利用した話題抽出手法に関してもさまざまな関連研究がある。戸田ら［14］は，Web ページ集合間の類似度をベースにグラフを構築し，Web ページの話題関連度・話題の重要度をノードの中心性により算出することで，話題の中心となる Web ページを発見する手法について提案している。Wang ら［15］はパイチャート及び線グラフといったシンプルなグラフを用いて製品と消費者の評価語を可視化し，製品に関する評判を抽出する手法を提案している。また Iino ら［16］は特許文書の集合を対象として，語の共起関係に基づくコンセプトグラフ（文書の階層関係を表現するグラフ）を作成し，組織が変化することでグラフ構造に変化がおきることを示した。

本研究でもグラフ構造を利用して評判を抽出するため，これらの既存研究とアプローチは似ている。それに加えて，我々の提案手法はグラフ構造の時系列変化を算出することで新たに発生した評判や勢いのある評判を発見する。グラフ構造の時系列変化は，単純な構造変化量ではなく，グラフの順序構造（階層）を考慮した変化量に基づいて算出することを目指す。この点が従来研究と大きくことなる点であると考える。

我々の手法は従来研究と異なり，予め対象とする製品を規定せず，不特定の製品に対して，想定外の消費行動が現れる可能性を考慮している。従来のマーケティング解析手法にはない新しいマーケティング解析手法であると言える。

４　時事問題をきっかけとした想定外の消費行動抽出手法

我々の手法は以下の７つのステップから構成される（図１）。

ステップ１：データクローリング
ステップ２：言語処理
ステップ３：共起抽出
ステップ４：グラフ生成
ステップ５：グラフ可視化
ステップ６：グラフ編集距離算出
ステップ７：消費行動抽出

以下，各ステップについて，簡単に説明する。

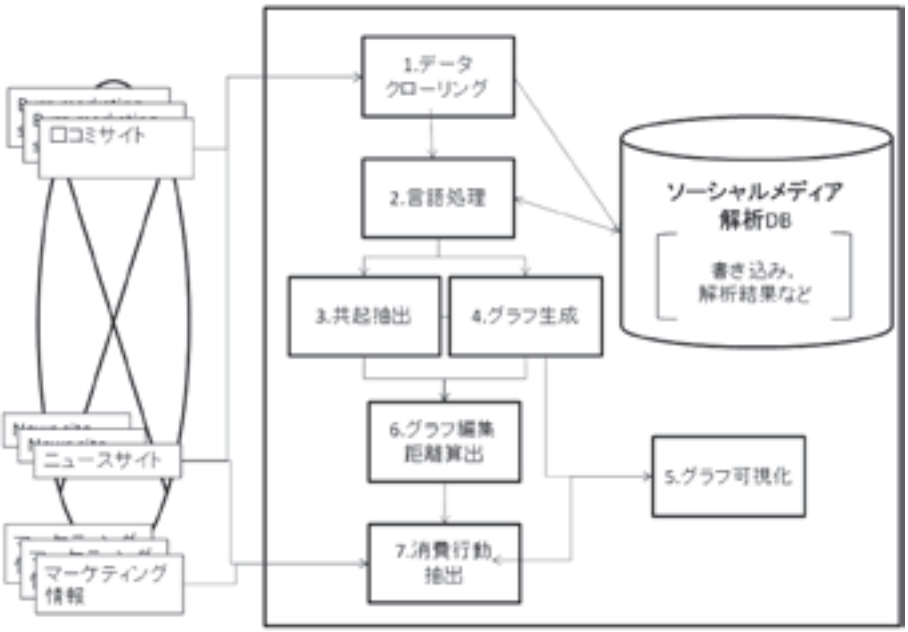

図１　提案手法のフレームワーク

4.1　ステップ１：データクローリング

本ステップは時事問題を表現する単語（「インフル」，「震災」など）を入力クエリとして，ソーシャルメディアの書き込みをクロールする。ソーシャルメディアにはブログや Twitter などさまざまな種類があるが，消費行動分析をターゲットとした本研究では，日本最大の製品価格比較サイトである「価格 .com」[17] を対象とする。価格 .com は製品別に口コミ掲示版を提供しており，消費者はその掲示板でコミュニケーションを取りながら，製品購入のための情報収集や購入後の感想などを述べている。我々はこの価格 .com の口コミ掲示版の書き込みから，消費者の消費行動を解析可能であると考えている。

図２はデータローリングの結果である。本ステップでは，書き込み ID，書き込み日付，日時，製品カテゴリ名，タイトル，書き込み者 ID，書き込み内容を収集している。ここで製品カテ

ゴリ名とは，「デジタルカメラ」や「空気清浄機」といった製品の総称を示している。今回の提案手法は，特定の製品を対象としておらず，製品全体の集合を対象としている。そのためクロールするデータは具体的な製品の名前ではなく，製品のカテゴリ名とした。クロール結果は消費者興味解析 DB に蓄積される。

4.2　ステップ２：言語処理

本ステップの入力は，ステップ1のデータクローリング結果である。ユーザの書き込み１件１件を１ドキュメントとしてみなし，形態素解析により名詞，動詞，形容詞，副詞を抽出する。さらに各単語のスコアを計算する。スコアリング手法としては，RIDF（residual IDF），LSA（latent semantic analysis），tf-idf（Term Frequency-Inverse Document Frequency）などが考えられるが，本研究では，内容語の抽出に適していると言われる RIDF を利用している。本ステップの出力は，書き込み ID，日付，製品カテゴリ名と，抽出された特徴語のリストである。この出力は，ステップ1と同様に消費者興味解析 DB に蓄積される。

4.3　ステップ３：共起抽出

本ステップでは，時事問題を表現する単語（「インフル」など）と製品間の共起関係を評価し，時事問題に影響を受けていそうな製品カテゴリ候補を見つける。入力はステップ２の言語処理結果である。時事問題を表現する単語と製品名の頻度情報をカウントし，時系列データとして共起しているかを判別する。ある時事問題に対して，共起していそうな製品カテゴリがあったとき，その製品カテゴリに対して消費者が何がしかの消費行動を行っている可能性があると予想する。共起関係の抽出には，Dynamic Time Warping（DTW）法［９］を利用する。以下，DTW 法について簡単に説明する。

4.3.1　Dynamic Time Warping（DTW）法

DTW 法は，２つの時系列データ間の類似性を評価する手法である。タイムスケールが異なっていても，時系列の形状が似ていれば類似性を評価することができる。DTW 法は音声認識などに利用されているが，我々はこの手法を時事問題と製品カテゴリの相関を判断するために用いている。時事問題をきっかけとして，製品カテゴリが何がしかの影響を受けるときに，たとえタイムラグがあったとしてもその類似性を評価できることから，この手法を利用することとした。

DTW 法の基本的な定義は以下のとおりである。

- 定義１．２つの時系列データ ts_1, ts_2 のローカルコスト行列 $C \in R^{|ts_1| \times |ts_2|}$

$$C_{i,j} \in \| ts_1[i] - ts_2[j] \|,\ i \in \langle 1 \ldots |ts_1| \rangle,\ j \in \langle 1 \ldots |ts_2| \rangle \quad (1)$$

ここで $\| ts_1[i] - ts_2[j] \|$ は時系列データ間の２点間の距離を示す。

このコスト行列により，DTW は２つの時系列データ間のコストを最小化するアラインメイトパスを生成する。このアラインメントパス p を “warping path” と呼び，以下のように定義していく；

- 定義２．時系列データの各点のペア

$$Pair_l = (pair_1, \ldots pair_k) \quad (2)$$

ここで $Pair_l = (i, j) \in \langle 1 \ldots |ts_1| \rangle \times \langle 1 \ldots |ts_2| \rangle$ は，２つの時系列データ ts_1, ts_2 の各点のペアを示す。

各ペアは，ts_1, ts_2 の時系列オーダを保持しており，warping path の最初と最後の点が各時系列データの最初と最後の点に一致する。warping path のコストは以下のように産出される；

- 定義３．warping path p のコスト

$$c(p)=\sum_{i=1}^{k} c(pair_1) \qquad (3)$$

DTW はコストを最小化した warping path として以下のように定義される；

- 定義４．２つの時系列データ間の *DTW*

$$DTW(ts_1,ts_2)=min\left\{c(p)|\ p\in P^{|ts_1|\times|ts_2|}\right\} \qquad (4)$$

ここで P はすべてのとりうる warping path を示し，その中で最小コストの warping path が DTW となる。

4.4 ステップ４：グラフ生成

本ステップは，ステップ２の特徴語抽出結果を入力として，ステップ３と平行して実施される。消費者の関心の推移を表現するために，有効グラフを生成する。有効グラフとしては廣川ら［18］が提案したコンセプトグラフを用いる。

コンセプトグラフは，ドキュメント集合に現れる単語の共起関係と頻度情報に基づいて，語の上下関係を動的に算出し，親に当たるノードを左側に，子に当たるノードを右側においた有向グラフを生成する手法である。たとえば「インフル」が入力クエリ（時事問題を表現する単語）だった場合,「インフル」をルートとして，それに共起する下位の単語を下位のノードとし，エッジで接続する。すべての上位語にエッジを描くのではなく，隣接上位にのみエッジを描く。

エッジには，その両端のノードの共起関係を保有する書き込みの製品カテゴリ名がラベルとして付加される。ある製品カテゴリにおいて，同様の共起関係をもつ書き込みが複数存在する場合は，書き込み件数の合計が，そのエッジの重みとなる。

以下，コンセプトグラフについて簡単に説明する。

4.4.1 コンセプトグラフ

検索対象全体のドキュメント集合を U とする。U の部分集合を X とし，|X| は X に含まれる文書の個数を表す。単語 u，v について，df(u,X) は u を含む X 中の文書数，df(u*v, X) は u と v の両方を含む X 中の文書数を示す。このとき，単語の関係を以下のように定義する。

$$r\ (v,u)\ =df(u*v,X)\ /\ df(v,X) \qquad (5)$$

ここで，$r\ (v,u)>0.5$ かつ $df\ (u,X)>df\ (v,X)$ ならば 単語 u は単語 v の上位にあると考える。具体的な可視化においては，単語 v についてすべての上位語に枝を描くのではなく，その隣接上位だけとすることで枝の数を抑えている。

4.5 ステップ５：グラフ可視化

本ステップは，ステップ４のグラフ生成において生成されたコンセプトグラフデータを可視化する処理部である。前節でも述べたように，コンセプトグラフの可視化においては，すべての上位語に枝を描くのではなく，隣接上位にのみ枝を描くため，比較的シンプルな有効グラフ

を得ることができる。

4.6　ステップ６：グラフ編集距離算出

我々の仮説は，コンセプトグラフの構造で大きな変化があったときに，消費者の行動にも変化があったのではないかということである。消費行動の変化を検知するために，本提案では時系列でグラフ・トポロジーの距離変化を測るグラフ編集距離［19］を指標として導入している。グラフ編集距離の時系列変化を算出することで，消費者の興味の推移や行動の変化を表現することができると考える。以下，グラフ編集距離について説明する。

4.6.1　グラフ編集距離

コンセプトグラフは，G =（V, E, α , β）として表現できる。ここで V は該当するノード集合であり，$E \subseteq V \times V$ はエッジの集合である。各ノードはラベル関数 $\alpha : V \rightarrow Lv$ によりラベル付けされている。ここで Lv はノードラベルの集合である。またエッジはラベル関数 $E{:}V \rightarrow L_E$ によりラベル付けされている。ここで L_E はエッジラベルの集合である。

前述のように，我々はコンセプトグラフのトポロジーの時系列変化を評価するためにグラフ編集距離を利用している。編集距離とは，あるグラフが別の構造を持つグラフに遷移するときの編集コストを表現している。一般にグラフ編集距離の計算は，編集の負荷が大きいが，幸いなことに今回我々が生成したコンセプトグラフのラベルは一意である。そのため，２つのグラフ（$G1 = (V_1, E_1, \alpha_1, \beta_1)$ 及び $G1 = (V_2, E_2, \alpha_2, \beta_2)$）間のグラフ編集距離は以下のように表現できる。

$$D_e(G_1, G_2) = |V_1| + |V_2| - 2\left|\alpha(V_1) \cap \alpha(V_2)\right| + |E_1| + |E_2| - 2\left|\beta(E_1) \cap \beta(E_2)\right|$$

ここで |G| はグラフ G のサイズ（グラフ G のエッジの数）を表現する。また α（V）は $\{\alpha(e) \in Lv \mid v \in E\}$ として定義される。同様に β（E）を $\{\beta(e) \in Le \mid e \in E\}$ として定義する。グラフ編集距離はグラフ構造の時系列変化を表現する指標であるため，グラフ編集距離により構造の変化を検知できる。

グラフ編集距離の算出は，ステップ３で評価した製品カテゴリ候補を中心に行う。コンセプトグラフ構造上で，エッジラベルの値が製品カテゴリ候補であるようなサブグラフを探索し，サブグラフの構造変化を編集距離により求めていく。

4.7　ステップ７：消費行動抽出

ステップ６で算出したグラフ編集距離に基づいて，グラフ構造が大きく変化した点（時系列上の時期）に消費行動の変化があったと想定する。製品の出荷台数といったようなマーケティングデータと照らし合わせ，グラフ構造の変化との相関を確認する。消費行動の変化が時事問題と一見関係のないように思える製品がテゴリについて起きていたとき，それを想定外の消費行動として評価する。

5　実験結果

提案手法に基づき，実際のデータを用いて実験を行った。本章ではその結果について述べていく。

前述のように本実験では，ソーシャルメディアとして，日本最大の製品価格比較サイトである「価格 .com」[17] を対象としている。

時事問題としては2009年に発生した新型インフルエンザの流行を取り上げる。2009年に流行した新型インフルエンザは，日本国内のみならず世界的にも大きな社会問題となった。消費者の消費行動にも大きな影響を与えたと考える。

以下，各ステップにおける処理結果を示す。

5.1 ステップ1：データクローリング

時事問題を表現する単語（検索クエリ）を「インフル」とし，2009年1月～12月までの口コミのクロールを行った。結果として，857件の書き込みが収集された。図2はクロール結果の抜粋である。なお書き込み ID，書き込み者 ID 情報等は伏字としている。

書き込み ID	書き込み日付	製品カテゴリ	書き込みタイトル	書き込み者	内容
******	2009/5/2	デジタルカメラ	豚インフルエンザ	****	今年は豚インフルがはやって海外旅行のキャンセルが相次いだため...
******	2009/5/2	ニュース総合	豚インフルエンザ	****	インフルエンザや花粉を完全防御するためにフルフェイス型ヘルメットを着用する人が...
******	2009/5/2	デジタルカメラ	We Love..	****	こちらでは観光産業が打撃を受けています。関西からの修学旅行キャンセルが相次いでいると，昨日行ったホテルのマネージャーが言っていました。...
******	2009/5/3	ニュース総合	豚インフルエンザ	****	インフルエンザが終息するまで我慢ですね。
******	2009/5/4	デジタルカメラ	息子の運動会	****	インフルエンザのおかげで息子の運動会がキャンセルになってしまい...
******	2009/5/4	デジタルカメラ	豚インフルエンザ	****	インフルが怖くて，結局旅行を取りやめました...
******	2009/5/15	空気清浄機	豚インフルエンザ	****	季節性インフルウイルスや，鳥インフルウイルス H5N1 ではテスト済み。豚由来のインフルエンザのウイルス H1N1は...
******	2009/5/4	空気清浄機	豚インフルエンザ	****	空気清浄機購入のきっかけになったのは「豚インフル」でした

図2　データクローリングの結果

5.2 ステップ2：言語処理

ステップ1のクローリング結果を入力とし，1件の書き込みを1ドキュメントとみなして，

特徴語抽出を行う。単語のスコアリング手法はRIDFを用い，特徴語抽出の閾値$T=1.0$とした。図３はステップ２の結果，抽出された特徴語のリストである。

書き込みID	書き込み日付	製品カテゴリ	重要語
******	2009/5/2	デジタルカメラ	インフル，旅行，キャンセル,...
******	2009/5/2	ニュース総合	インフル，花粉，マスク,...
******	2009/5/2	デジタルカメラ	インフル，観光，キャンセル，関西,...
******	2009/5/3	ニュース総合	インフル，終息，我慢,...
******	2009/5/4	デジタルカメラ	インフル，運動会，キャンセル,...
******	2009/5/4	デジタルカメラ	インフル，旅行，取り，止め,...
******	2009/5/15	空気清浄機	インフル，空気，清浄，ウイルス，予防,...
******	2009/5/4	空気清浄機	インフル，空気，清浄，購入,...

図３　言語処理による特徴語抽出の結果

5.3　ステップ３：共起抽出

ステップ２の結果を入力として，インフルと製品カテゴリ名の時系列の共起関係をDTW法により評価する。DTW距離の算出にはフリーの統計解析ソフトR［20］を用いる。本ステップでは，価格.comの製品カテゴリ名とインフルの頻度情報に基づく共起関係を算出している。時系列に比較を行うため，書き込みデータを月ごとなどにグループ化し，頻度情報をカウントし，時系列データとして，DTW法により距離を算出した。算出結果のうち，ここでは，カメラ，空気清浄機，車，プリンター，携帯電話，テレビの結果について説明する。

図４はインフルと空気清浄機，カメラの時系列頻度推移，図５はインフルと車，プリンター，携帯，テレビの時系列頻度推移である。図４を見てもわかるように，インフルと空気清浄機，カメラの時系列頻度推移は相関していると予想できる。一方，インフルと車，プリンター，携帯，テレビの時系列頻度推移は，空気清浄機やカメラと比較すると相関が弱いように思われる（図５）。図６－11はインフルと上記６つの製品の最小コストのwarping pathである。図６がインフルと空気清浄機，図７がインフルとカメラ，図８がインフルと車，図９がインフルとプリンター，図10がインフルと携帯，図11がインフルとテレビのwarping pathを示している。

図６のインフルとカメラ，図７のインフルと空気清浄機のwarping path, は他の組み合わせに比べると短い。これはインフルとカメラ，インフルと空気清浄機が共起関係にあることを示している。

表１は，インフルと６つの製品カテゴリのDTW距離を示している。上記でも述べたようにインフルとカメラ，インフルと空気清浄機のDTW距離は短い。一方それに比較して，インフルと車，インフルとプリンター，インフルと携帯，インフルとテレビ間のDTW距離は長い。カメラと空気清浄機はインフルと相関があると考えられる。インフルをきっかけとして何らかの消費行動が起きていると想定される製品群（製品候補）とみなされ，本ステップの出力結果となる。

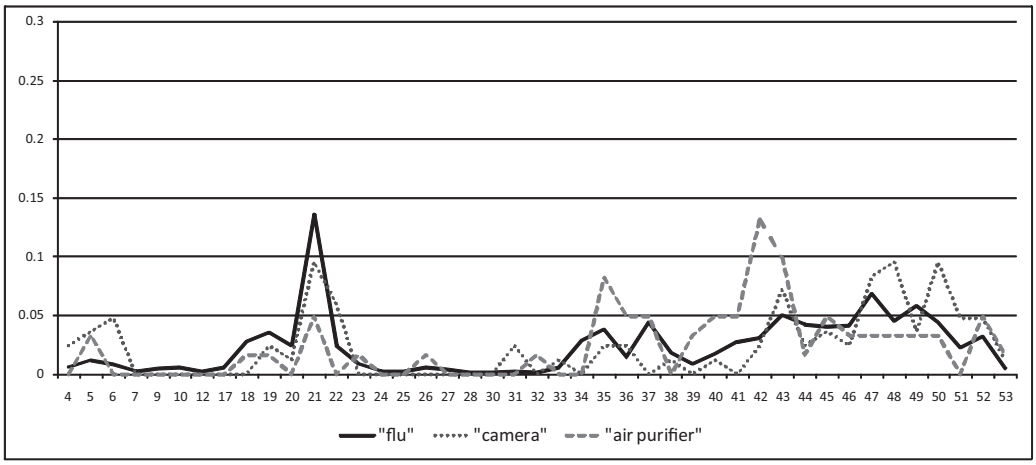

図４ インフルと空気清浄機，カメラの単語頻度の時系列頻度

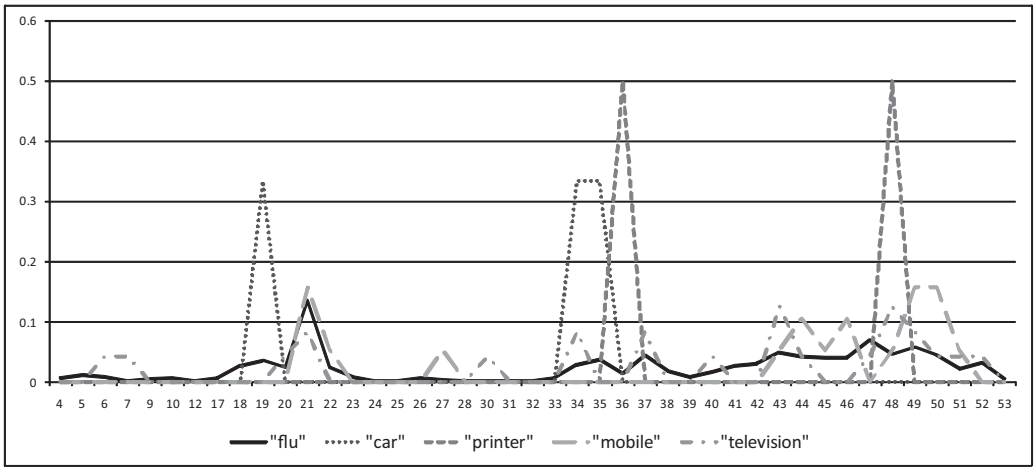

図５インフルと車，プリンター。携帯，テレビの単語頻度の時系列頻度

表１　インフルと代表的な製品カテゴリ間の DTW 距離

	カメラ	空気清浄機	車	プリンター	携帯	テレビ
インフル	0.821	0.870	1.816	1.948	1.049	1.027

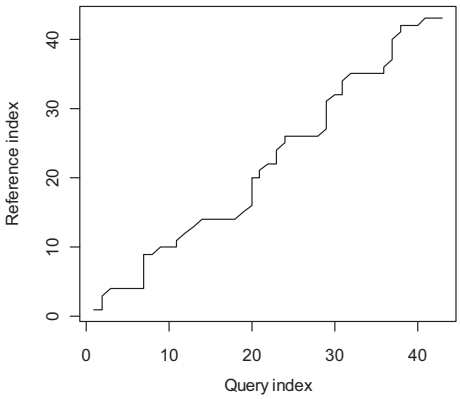

Fig. 6.インフルとカメラのDTW

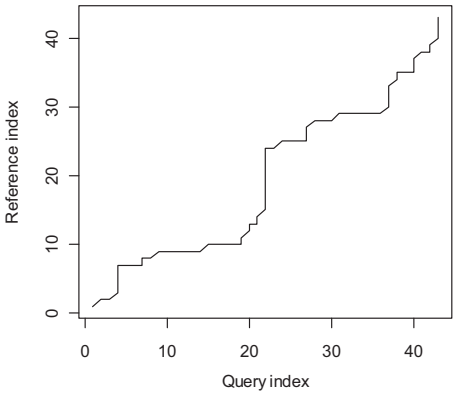

Fig. 7.インフルと空気清浄機野DTW

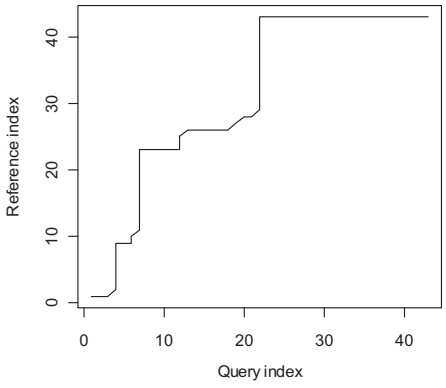

Fig. 8.インフルと車のDTW

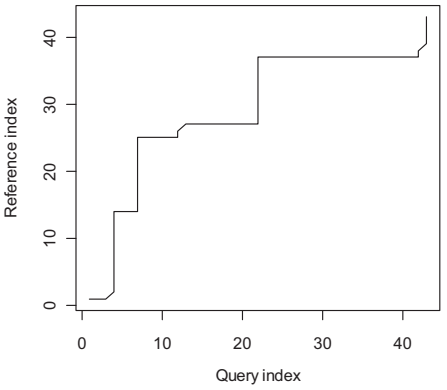

Fig. 9.インフルとプリンタのDTW

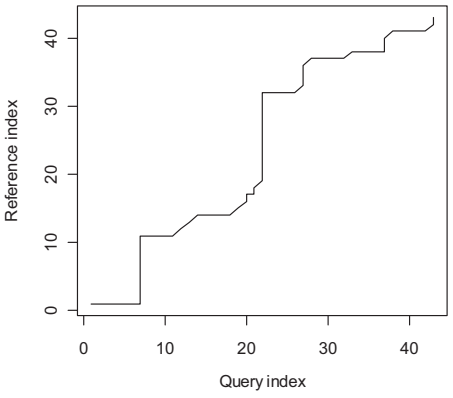

Fig. 10.インフルと携帯のDTW

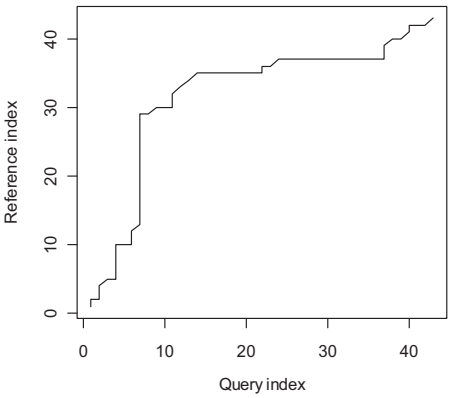

Fig. 11.インフルとテレビのDTW

5.4　ステップ4：グラフ生成

本ステップは，ステップ２の結果を入力として，ステップ３と平行して実施される。消費者の消費行動推移を表現するために，コンセプトグラフにより有効グラフを生成する。コンセプトグラフのデータは Gexf（Graph Exchange XML Format）形式で保存される。Gexf は，複合的ネットワーク構造を表現するための言語であり，階層構造を持ち，エッジにラベル・重みを持つ今回のコンセプトグラフの表現には最適な言語であると考える。

5.5　ステップ５：グラフ可視化

ステップ４のアウトプットである Gexf ファイルを入力として，有効グラフを可視化する。グラフの可視化はネットワーク・動的グラフ及び階層グラフのためのインタラクティブな可視化・探索プラットフォームである Gephi［21］を利用する。図12は Gephi による可視化の結果である。時系列を表現するスライダを動かすことにより，時間的な変化を評価することができる。

また図13はコンセプトグラフを月ごとに時系列で可視化したものである。図13のグラフ構造の中に，空気清浄機に関するサブグラフとデジタルカメラに関するサブグラフが見られ，それらが時間が経つにつれ構造変化している様子が見て取れる。実際にデジタルカメラのサブグラフにおいては，インフルに共起した単語として旅行，キャンセル，運動会，購入といった単語が並んでいたことが我々の観察でもわかっている。インフルエンザが，デジタルカメラに関心のある消費者に何らかの影響を及ぼしていることが予想される。

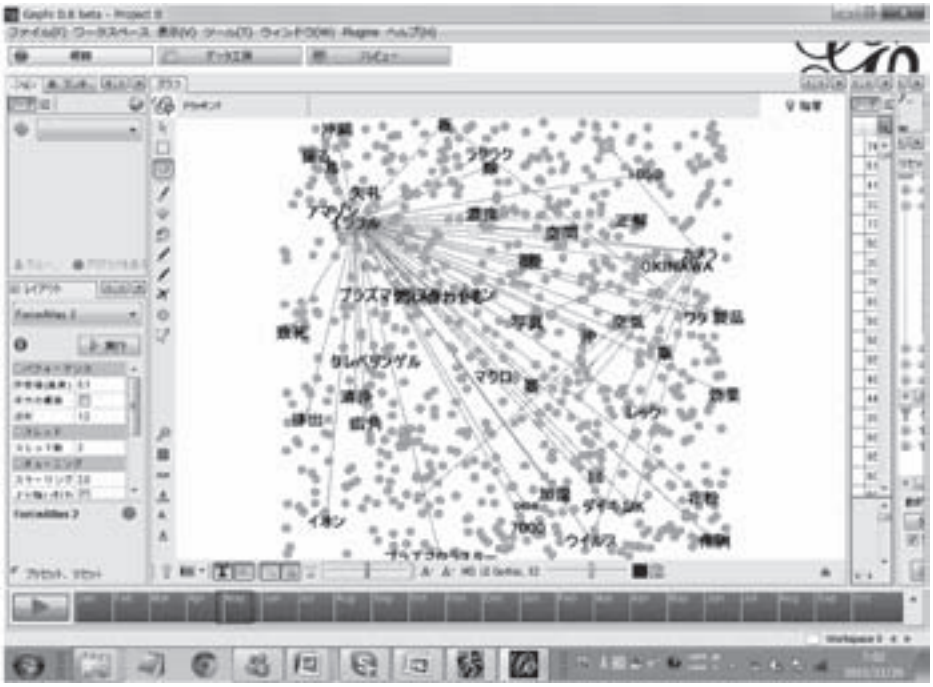

図12　Gephi による可視化例

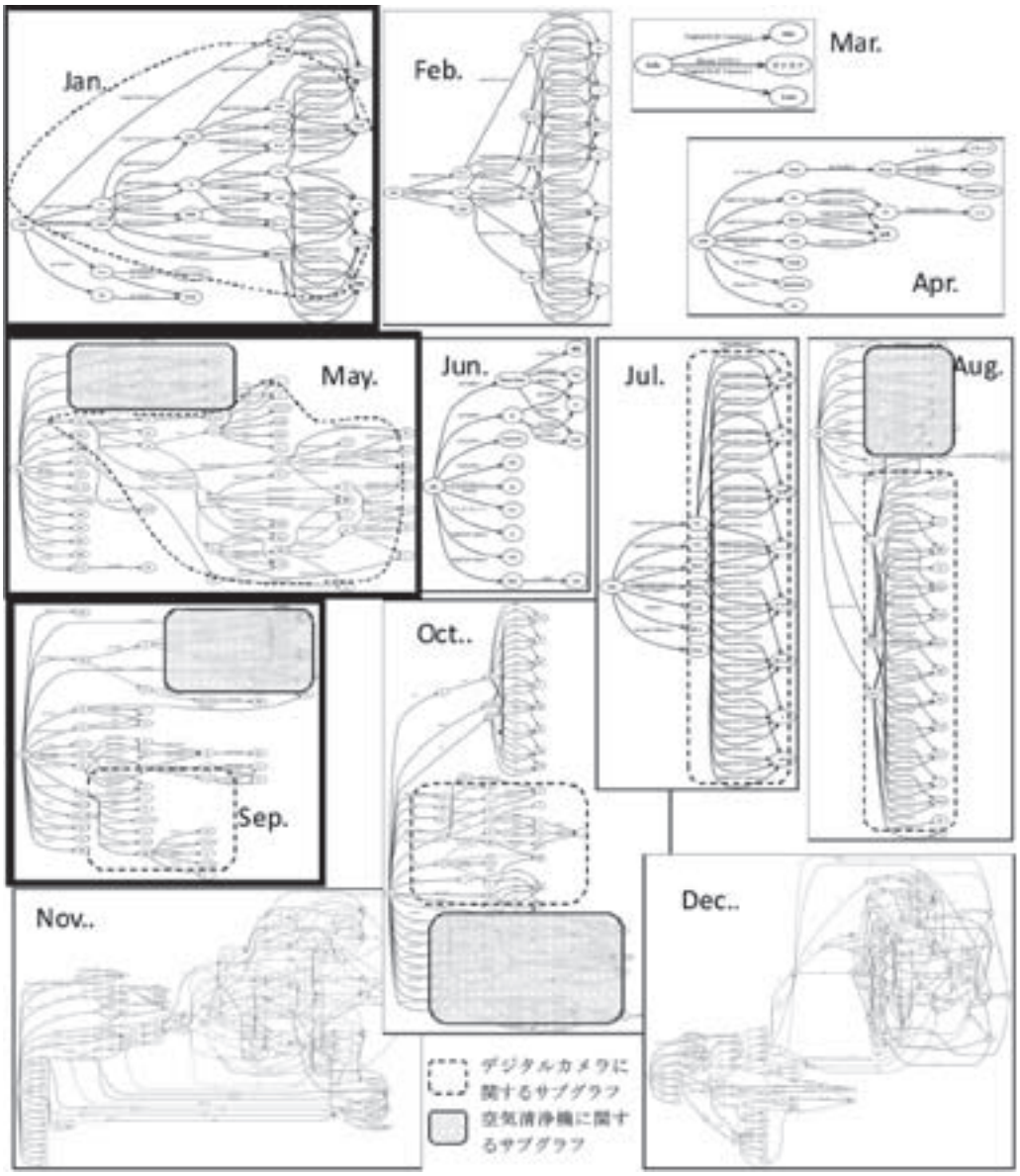

図13　月別コンセプトグラフの時系列変化

5.6　ステップ６：グラフ編集距離算出

ステップ３の結果である製品候補に対して，ステップ４で算出したグラフ構造に対して，グラフ編集距離を算出する。ステップ４で算出したグラフ構造は時系列情報である。ステップ３の各製品候補をラベルとして含むサブグラフが月別のグラフに存在するか否かをチェックし，製品候補ごとにグラフ編集距離を算出する。図14はデジタルカメラと空気清浄機に関して編集

距離を算出した結果である。デジタルカメラは５月と10月に，空気清浄機は５月と９月に大きな構造変化があることがわかる。

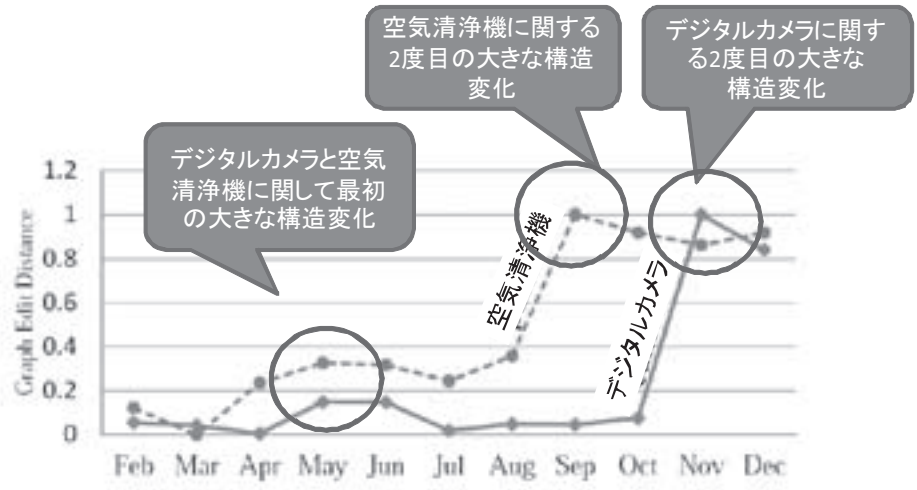

図14　デジタルカメラと空気清浄機のグラフ編集距離算出結果

5.7　ステップ７：消費行動抽出

図14のグラフ編集距離の結果では，空気清浄機に関する構造の編集距離は５月と９月に大きく変化していることがわかる。図15において，空気清浄機の構造は灰色の矩形で表現されているが，特に５月と９月に大きな構造変化が見られていることがわかる。実際に2009年５月には新型インフルエンザの世界的な流行が始まり，日本においても最初の患者が発見されている。消費者はこうしたことに敏感になり，５月頃に空気清浄機の口コミサイトにおいて，インフルエンザの話題が活発化したと考えられる。

さらに2009年秋には，新型インフルエンザは日本にで本格的に流行し，結果として，ウイルス除去機能を持つ空気清浄機に対する関心がますます高まったのではないかと予想される。図15は2009年の空気清浄機の出荷台数である。昨年に比べ４，５月と９，11月に大きく出荷台数が増加している。

インフルエンザの流行に反応して，空気清浄機を購買する，といった行動は想定内の消費行動である。こうした想定内の消費行動も我々の手法により発見することができる。

一方，図14でデジタルカメラは５月と10月に大きな構造変化が見られる。また図13でも，デジタルカメラのサブグラフ構造は点線の矩形で表現されているが，特に５月と９，10月に大きな構造変化が見られたことがわかる。図16は2008年と2009年のデジタルカメラの出荷台数である。2008年の出荷台数は点線で，2009年の出荷台数は実線で表現されている。2009年の５月及び，９，11月の出荷台数が2008年に比べて減少していることがわかる。これは，グラフ編集距離の構造変化に対してデジタルカメラの消費行動がネガティブに反応していることを示すものである。コンセプトグラフのノードには，旅行，運動会，キャンセルといった単語が提示され

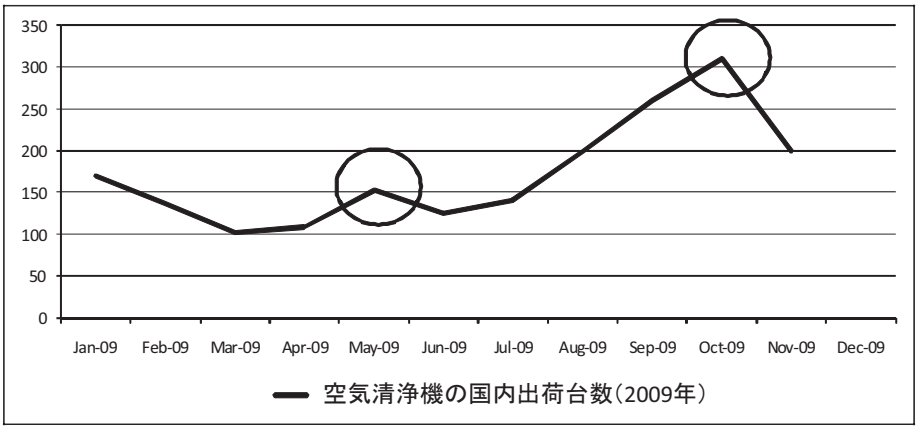

図15　空気清浄機の国内出荷台数（2009年）

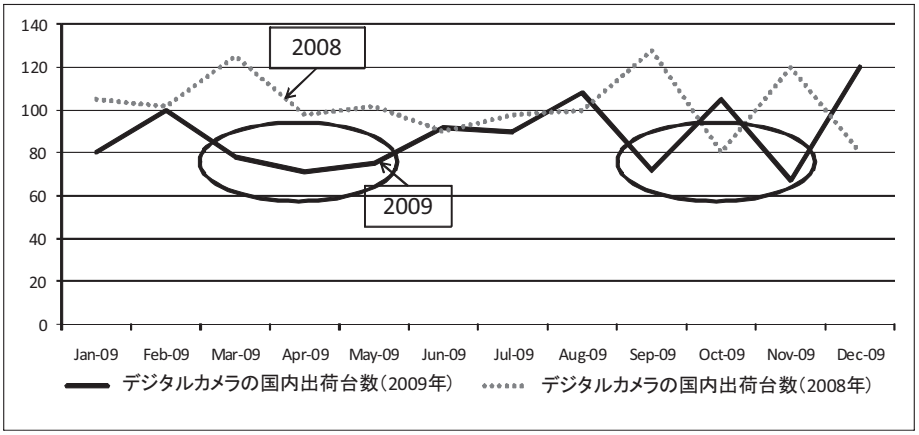

図16　デジタルカメラの国内出荷台数（2008年と2009年）

ており，インフルエンザの流行によってキャンセルされたイベントが，デジタルカメラの購入を控えるという消費行動を呼び起こしているのではないかと予想される。こうした時事問題に一見無関係に思える製品カテゴリを対象とした消費行動を我々は想定外の消費行動と呼んでいる。我々の手法により，このような想定外の消費行動を発見することが可能となる。

6　結論

本稿では，ソーシャルメディアを対象として，時事問題をきっかけとした想定外の消費行動抽出手法について提案した。我々の手法は時事問題と各種製品間の時系列相関を算出し，時事問題との間に想定外の相関関係をもつような製品カテゴリ候補を抽出する。さらに語の共起関係をベースに口コミサイトの書き込みを構造化し，話題構造の推移を時系列で可視化する。時系列グラフ構造の動的な振舞いを分析することで，時事問題をきっかけとした想定外の消費行動を抽出する。我々の手法により，従来は抽出が難しかった時事問題と一見無関係に思える製

品に対する消費者の行動を分析することが可能となった。

今後はさらに多くのデータに本手法を適用し，提案手法の効果を明らかにしていくとともに，東日本大震災と言った災害におけるソーシャルメディアの口コミにも適用し，消費行動だけでない評判や風評の解析にも展開していく予定である。

謝辞

本研究の一部は学習院大学計算機センター特別研究プロジェクト「概念グラフを用いた化粧品のニーズ分析」により支援されました。ここに記して謝意を表します。

References

1. Nagano, S., Inaba, M., Mizoguchi, Y., Iida, T., Kawamura, T.: Ontology-Based Topic Extraction Service from Weblogs. IEEE International Conference on Semantic Computing, pp.468-475, 2008
2. Kobayshi, N., Inui, K., Matusmoto, Y., Tateishi, K., Fukushima,, S.; Collecting evaluative expressions by a text mining technique, IPSJ SIG NOTE, Vol.154, No.12, pp. 77-84, 2003.
3. Asano, H., Hirano, T., Kobayasi, N., Matsuno, Y.: Subjective Information Indexing Technology Analyzing Word-of-mouth Content on the Web, NTT Technical Review, Vol. 6 No. 9 Sep. 2008, pp.1-7, 2008
4. Spangler, W.S., Chen, Y., Proctor, L., Lelescu, A., Behal, A., He, B., Griffin, T.D., Liu, A., Wade, B., Davis T.: COBRA - mining web for COrporate Brand and Reputation Analysis. Web Intelligence and Agent Systems（WIAS）7（3）, pp.243-254, 2009.
5. Wang, G., Araki, K.: A Graphic Reputation Analysis System for Mining Japanese Weblog Based on both Unstructured and Structured Information, AINA Workshops 2008, pp.1240-1245, 2008
6. Hashimoto, T., Shirota Y.: Semantics Extraction from Social Computing: A Framework of Reputation Analysis on Buzz Marketing Sites, Lecture Notes in Computer Science, 2010, Volume 5999/2010, pp.244-255, 2010
7. Hashimoto T., Kuboyama T., Shirota Y.: Graph-based Consumer Behavior Analysis from Buzz Marketing Sites, Proc. of 21st European Japanese Conference on Information Modelling and Knowledge Bases, pp.60-71, 2011.
8. Kuboyama T., Hashimoto T., Shirota Y.: Consumer Behavior Analysis from Buzz Marketing Sites over Time Series Concept Graphs, Proc. of 15th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, pp. 73-83, 2011.
9. Berndt D., Clifford, J.: Using dynamic time warping to find patterns in time series, Proc. of Advances in Knowledge Discovery and Data Mining, pp. 229-248. AAAI/MIT, 1994.
10. Hashimoto,, T., Kuboyama T., Shirota Y.: Detecting Unexpected Correlation between a Current Topic and Products from Buzz Marketing Sites, Proc. of the DNIS 2011, LNCS 7108, 2011（to appear）.
11. Zhu Y., Shasha D.: Warping Indexes with Envelope Transforms for Query by Humming, Proceedings of the 2003 ACM SIGMOD international conference on Management of data, pp. 181-192, 2003.
12. Otranto E.: Identifying financial time series with similar dynamic conditional correlation, Journal of Computational Statistics & Data Analysis archive Volume 54 Issue 1, pp.1-15, January, 2010.
13. C. Loy, T. Xiangand S. Gong, “Time-Delayed Correlation Analysis for Multi-Camera Activity

Understanding", International Journal of Computer Vision（IJCV）, vol. 90, no. 1, pp.106-129, October 2010.
14. 戸田 , 北川 , 藤村 , 片岡 , 奥 : グラフ分析を利用した文書集合からの話題構造マイニング, 電子情報通信学会論文誌, Vol. J90-D, No. 2, pp.292-310, 2007年２月
15. Wang, G., Araki, K.: A Graphic Reputation Analysis System for Mining Japanese Weblog Based on both Unstructured and Structured Information, AINA Workshops 2008, pp.1240-1245（2008）.
16. Iino, Y., Hirokawa, S.: Time Series Analysis of R&D Team Using Patent Information, Lecture Notes in Computer Science, 2009, Volume 5712/2009, pp.464-471（2009）.
17. kakaku.com, http://kakaku.com/
18. Shimoji, Y., Wada, T., Hirokawa, S.: Dynamic Thesaurus Construction from English-Japanese Dictionary, The Second International Conference on Complex, Intelligent and Software Intensive Systems, pp.918-923, 2008.
19. Bunke, H.: On a relation between graph edit distance and maximum common subgraph, Pattern Recognition Letters, Volume 18, Issue 8, August 1997, pp.689-694, 1997
20. R, a language and environment for statistical computing and graphics, http://www.r-project.org/
21. Gephi. http://oss.infoscience.co.jp/gephi/gephi.org/index.html