Vol.2014-IFAT-114 No.10
Vol.2014-DD-93 No.10
2014/3/29

EC サイトにおける購買行動促進のための重要語抽出とタグクラウド生成

松崎 友見[1,a)]　波多野 賢治[1,b)]

概要：インターネットショッピングサイト (以下，EC サイト) は，時間帯や場所にとらわれず商品の購買ができるため，現在市場規模の発展が目覚ましいコンテンツである．しかしユーザ数の増加率に比べると，購買人数の増加率は少ない．これは，EC サイトのユーザが商品の購買判断をする際，商品に関するデータ量が過多であり商品の特徴を素早く捉えることができないため，購買行動に結びつきにくいことが原因であると考えられる．そのため，購買判断に重要な役割を果たすレビューから重要語を抽出することで，購買人数の増加，更に通信販売業の発展や活性化にも繋がると考えられる．本稿では，係り受け解析器，日本語概念辞書，データ視覚化技術を用いて商品紹介やレビューから重要語抽出し，それらをタグクラウドを用いて提示する方法を提案する．また，提案した方法により商品の購買行動につながるかどうかを評価実験において確認する．

1. はじめに

現在，Amazon.com や楽天市場といった EC サイト の利用者は年々増加傾向にあることが，経済産業省によって実施された調査により報告されている [*1]．これは，消費者の商品購買が実店舗に限らず EC サイトをも利用する形態に変化し始めていることを示しており，商品の質や価格での差別化が難しくなってきた現在，消費者が商品を購入することだけに着目するのではなく，商品を利用するプロセスをも商品選択の際に考慮し始めていることを示している．実際，消費者が EC サイト上の商品レビューを参考に複数の商品を比較しながら商品購買を行っているという報告もある [*2] もある．

この傾向は，ソーシャルメディアの発達により消費者の購買経験が消費者自身により発信され，それらが多くの人々により容易に共有されるようになった時点から表面化し，それ以後，消費者は商品の良い点，悪い点を消費者の経験価値として分け隔てなく共有できるようになった．この情報共有の供給源は EC サイト内のレビューにあることから，EC サイトが年々成長し続けている経済市場である理由は，EC サイト内のレビューが消費者にとってある一定の価値を有しているからとも言える．しかし，消費者が購買判断の材料として使用するレビューのデータ量は過多であり，商品の特徴を消費者は素早く捉えることができない．そのため，レビューデータの存在だけでは消費者の購買行動に直接結びつかなくなってきている．そのため，購買判断に重要な役割を果たすレビューを整理し，購買判断時間の短縮を行うことで，購買人数の増加，更に通信販売業の発展や活性化にも繋げていく必要がある．

そこで本稿では，大量のデータを整理するための一手法であるソーシャルタギング [1] を利用してレビューデータを整理するために，係り受け解析器である KNP [2]，日本語 WordNet [3] を用いて機械的に商品紹介やそのレビューからタグにふさわしく表記揺れのない語を抽出し，最終的にデータ視覚化技術であるタグクラウドを用いて視覚化する方法を提案する．また，提案方法により消費者の商品購買行動を促せるか否かを評価実験において確認する．

2. 関連研究

本稿では EC サイト利用者の購買経験を整理，視覚化するためにタグクラウドを用いている．このとき，タグクラウドに表示される語の抽出は多数の EC サイト利用者によって書かれたレビューデータを利用するが，どの語を重要とし抽出するか，また表記に揺れのある語が指し示す対象を機械的に獲得することは難しい．

本節では，複数ユーザによって書かれた文書からの重要

[1] 同志社大学文化情報学部
〒 610–0394 京都府京田辺市多々羅都谷 1-3
[a)] bik136@mail4.doshisha.ac.jp
[b)] khatano@mail.doshisha.ac.jp

[*1] http://www.meti.go.jp/press/2013/09/20130927007/20130927007-4.pdf 2014 年 2 月閲覧

[*2] http://www.nielsen-online.com/pr/pr_081218.pdf 2014 年 2 月閲覧

語の抽出手法と表記の揺れに対する対処法，そして抽出した語の視覚化に用いるタグクラウドについて概説する．

2.1 談話的文書からの重要語句抽出

語に対する重要度決定の手法として，よく利用される手法に TF-IDF 法が挙げられる [4] ．TF-IDF 法はある N 個の文書群に着目し，出現する語句の頻度情報を基に各文書における語句の重要度を決定するものであり，文書 D_j 中に含まれる語 T_i の出現頻度を $n(T_i, D_j)$，語 T_i を含む文書数を $N(T_i)$ とした場合，文書 D_j 中に含まれる語 T_i の重要度 $\mathrm{TFIDF}(T_i, D_j)$ は以下のように計算される．

$$\mathrm{TF}(T_i, D_j) = \frac{n(T_i, D_j)}{\sum_i n(T_i, D_j)} \tag{1}$$

$$\mathrm{IDF}(T_i) = 1 + \log \frac{N}{N(T_i)} \tag{2}$$

$$\mathrm{TFIDF}(T_i, D_j) = \mathrm{TF}(T_i, D_j) \cdot \mathrm{IDF}(T_i) \tag{3}$$

しかし日本語で TF-IDF 法を用いる場合は主語が頻繁に省略されるため，語の重要度を元に重要語抽出を行ってしまうと適切な語を重要語として抽出できない問題があることが指摘されている．

そのため飯田らは，TF-IDF 法のような単純に文書中の出現頻度から重要度を求めるのではなく，談話の顕現性に基づく語の集約を行っている [5]．談話の顕現性とは，ある語が繰り返し談話の中で使われるとき，繰り返し回数を重ねるごとにその語は省略されるが，内容からははっきりとその語の存在が伝わる性質のことである．この手法では，名詞句を談話のもつ二種類の顕現性に基づいて主語と主題，それらを修飾する語，それ以外を区別した上で，照応解析を用いて頻繁に省略され話題の中心として重要な役割を担うゼロ代名詞を特定することができるようになり，語句の重要度の計算を正確に行うことができるようになっている．

2.2 ソーシャルタギングにおける語の表記

一般に，あるデータに対してある語を注釈として付与することをアノテーションというが，これを複数のユーザによって行うことを最近はソーシャルタギングと呼ばれている [1]．ユーザごとにそのデータに対する認識が異なるため，そのデータに対してソーシャルタギングを行った場合，付与された語にしばしば表記の揺れが生じる．そのため，タグとして用いられている語が指し示す対象を機械的に獲得することは二つの理由から難しいとされている [6]．

一つ目の理由は，アノテーションに用いられる語自体が持つ曖昧さである．たとえば，“bat” という言葉は，“コウモリ” と “野球のバット” の両方を指す多義語であり，“bat” という言葉だけではどちらを指しているのかがわからない．また同じ言葉であってもユーザによってその認識範囲が異なるため，たとえば “bear” という言葉の指示対象に “パンダ” を含めるか否かは，ユーザの “bear” に対する認識違いによるものである．

もう一つの理由は，ノイズタグの問題である．ソーシャルタギングの特徴として，各データに対しユーザが自由にタグを付与できる点が挙げられているが，しばしばそのデータとは無関係なタグが付与されていることがある．その要因としては，ユーザがいくつかのデータをまとめてアノテーションを行う際に生じるアノテーション対象の選択ミスやタグのスペルミスなどが挙げられる．

このようにソーシャルタギングには，語に表記の揺れが生じるという問題が存在する．語の表記に揺れが生じた場合，タグの量が増加し，タグが表す意味を認識することが困難になるため，語の表記の揺れを機械的に抑制する手法がいくつか考えられている．その抑制法の一つとして，共起情報と辞書の語釈文を用いた語の意味の違いを認識する手法が挙げられる [7]．従来のように共起情報を利用した統計的単語クラスタリングでは，共起情報が語の意味間類似度の推定にある程度有効に働くため，語の意味の違いを認識するには有効性があったが，類義語間の微妙な意味の違いまでは判別できないという問題があった．そのため，文献 [7] では，語釈文中の共通語に対する修飾語句の違いから，語の意味の違いを判別している．

2.3 タグクラウド

抽出した重要語をコンテンツにアノテーションし，それらを表示する方法の一つにタグクラウドがある．タグクラウドは，ブログ等の Web サイト上でよく使われるキーワードをタグとして用い，それらを視覚的に記述する表現方法である [8]．大規模なサービスでタグクラウドが導入されたのは，写真共有サイト Flickr [*3] が最初である．

タグクラウドに表示されるタグは，サイト内で使われる語や，ユーザが付与する語とさまざまな手法で選出されている．タグクラウドでのタグの表示順は，通常アルファベット順，あるいは五十音順 (漢字コード順) に並べられ，タグの重要度に応じて文字サイズや太さを変えることで，より重要なタグが目立つように設計されている．そのため，Web サイトの訪問者は，タグクラウドを見ることで，そのサイトからどのような情報が発信されているかを視覚的に知ることができる．また，タグをクリックすることで，Web サイトの中でそのタグに関連した情報を示すこともできる．図 1 にタグクラウドの例を示す．

3. 購買行動促進システム

本節では，2 節で説明した談話的文書からの重要語句抽出法を利用して，EC サイトのから発信されるデータから

[*3] http://www.flickr.com/ 2013 年 11 月閲覧

Vol.2014-IFAT-114 No.10
Vol.2014-DD-93 No.10
2014/3/29

図 1　タググラウド

消費者に商品の購買行動を促すための情報を，タグクラウドを用いて視覚化するシステムについて述べる．このシステムの開発目的は 1 節でも述べたとおり，EC サイトユーザの購買判断に重要な役割を果たす EC サイト内のデータを整理，購買判断時間の短縮を行うことで，購買人数の増加，更に通信販売業の発展や活性化にも繋げていくためである．

構築したシステムでは，まず EC サイトの商品紹介とレビュー内のテキストデータから，EC サイトユーザの商品購買につながる重要語を抽出した後，抽出された重要語をタグクラウドを用いて視覚化することで，商品のデータ整理を行う．アノテーションの対象となる商品ページを $P_i(i=1,2,3,\ldots)$，それに含まれている商品紹介ページを $Q_i^j(j=1,2,3,\ldots)$，商品のレビュー文書を $R_i^k(k=1,2,3,\ldots)$ としたとき，提案するタギングシステムの処理手順の概要 (概略図は図 2) は以下のとおりである．

(1) P_i に含まれる商品ページ Q_i^j，レビュー文書 R_i^k 中の各文内に省略されている主語・主題の補完
(2) 各文を形態素に分割し，形態素ごとに重要度を算出，そして重要語候補の抽出
(3) 重要語候補を日本語 WordNet を持ちいて概念別にグルーピング
(4) 各概念グループを代表する重要語 (タグ) の決定
(5) 各概念を表現するタグを用いてタグクラウドの作成

以下の節で手順 (1) から (5) までの処理を詳述する．

3.1　主語・主題の補完

2.1 節で説明したように，日本語の場合，語の重要度を求めるためには主語，主題を含んだ文でなくては正確な語の重みを計算することができない．厳密な文には必ず主語と述語が含まれているが，アノテーションに使用する商品紹介とレビュー文書の性質を考えると必ずしもそれらが厳密な日本語文法を用いて書かれているとは限らない．したがって，ゼロ代名詞を含む文でも主題や主語を抽出できるように各商品 P_i に含まれる各文書 Q_i^j，R_i^k に対して，ゼロ照応解析を行うことで省略されているゼロ代名詞の補完し，主語を含んだ文を構築していく．

ゼロ照応解析では，係り受け解析の結果である名詞格を用いて顕現性の高さを

$$ハ格 > ガ格 > ニ格 > ヲ格 > その他$$

の順で定義した上で，文中の係り受け関係から顕現性の高いゼロ代名詞を判断し，そのゼロ代名詞を主語と主題が省かれている語を補完する．本稿では，日本語構文・格解析システム KNP [*4] を利用し，日本語に対して照応解析を行う．KNP は日本語形態素解析システム JUMAN [*5] の解析結果である形態素列を入力とし，文節および基本句間の係り受け関係と格関係，そして照応関係を出力するが，これらは Web から自動構築した大規模格フレームに基づく確率的モデルにより決定している．また，KNP は格構造解析に語の基本形ではなく，活用形を用いることから談話的な文書に対しゼロ照応解析を行うことが可能で，他の機器に比べ精度が高いと言われている．

3.2　重要語候補の抽出

次に，3.1 節で省略された語を補完した文を形態素に分割するが， これには前述したように KNP で利用されている形態素解析器 JUMAN の出力を利用する．形態素解析器を用いれば各文は形態素に分割されるが，これらの形態素からタグになりうる重要語候補の語を抽出するのが本ステップである．

まず，形態素解析を行った結果である形態素列を，前から順に助詞のすぐ後ろの部分に “，(カンマ)” を補完し形態素列を区切ることにより句を作成する．商品紹介ページ Q_i^j とレビュー文書 R_i^k から作られた句をそれぞれ $C_t^{\alpha}(Q_i^j)$，$C_s^{\beta}(R_i^k)$ とした場合，それらに含まれている形態素の品詞を調べ，連体語と接続語，独立語以外の自立語であれば，それを重要語候補 $w_q(Q_i^j)$，$w_r(R_i^k)(q,r=1,2,3,\ldots)$ として抽出する．自立語を重要語候補とした理由は，自立語はその語のみで意味が理解できるためである．

3.3　重要語を概念別にグルーピング

3.2 節で抽出した重要語候補 $w_q(Q_i^j)$，$w_r(R_i^k)$ の中には，同じ意味を持っていても別の表記がされている可能性が残されている．これは，商品紹介ページやレビュー文書は複数人で書いているため，語の意味を統一的に扱えない可能性があるためである．この表記の揺れをなくすために，類義語が含まれていた場合は，語の意味がどれか一つの語に定まりその結果消費者が商品を探しやすくなるように，日本語 WordNet を用いて表記の揺れを解消する．

日本語 WordNet は，独立行政法人情報通信研究機構

[*4] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP 2013 年 8 月閲覧
[*5] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN 2013 年 11 月

Vol.2014-IFAT-114 No.10
Vol.2014-DD-93 No.10
2014/3/29

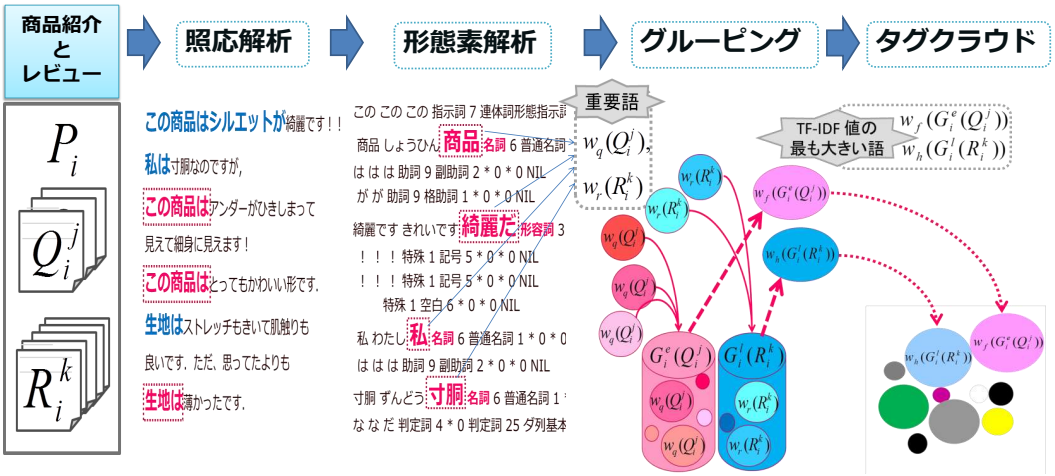

図 2　提案手法の流れ

(NICT) から公開されている日本語の意味辞書であり，一つの英単語を概念として日本語の同義語を集めた構造になっている [3]．57,238 種の概念それぞれに日本語の同義語が計 93,834 語付与されているため，ある概念とその概念に対する同義語の集まりを一つの概念グループとして扱うことで，同義語群を一つの概念にグルーピングすることができる．つまり，3.2 節で抽出された重要語候補 $w_q(Q_i^j)$，$w_r(R_i^k)$ を各概念にそれらが属しているか否かをチェックしそれらを各概念にグルーピングすることにより，表記の揺れを解消することができる．

同じ概念か否か判断するためには，日本語 WordNet の各語にに付与されている Synset 番号を用いればよい．Synset 番号は八桁の数字と品詞を表す一字のアルファベットから成り立っており，その上位五桁で概念関係が判別することができる [*6]．そのため，日本語 WordNet 用 Java API である JAWJAW [*7]．を用いて，抽出した各重要語候補に Synset 番号を付与し，上位五桁が一致した語を一つの概念グループを構成すると判断する．

以上の手順で，商品紹介の概念グループ $G_i^e(Q_i^j)(e = 1,2,3,\ldots)$ と，レビュー文書の概念グループ $G_i^l(R_i^k)(l = 1,2,3,\ldots)$ に重要語候補 $w_q(Q_i^j)$ と $w_r(R_i^k)$ を分類し，各概念ごとにグルーピングを行う．

3.4　タグの決定

3.3 節で重要語候補 $w_q(Q_i^j)$ と $w_r(R_i^k)$ は商品紹介とレビュー文書の概念グループ $G_i^e(Q_i^j)$，$G_i^l(R_i^k)$ に振り分けられたが，各概念グループ内の代表語が決まっていないため，そのグループを表現するタグが決まっていない．そのため，最も重要度が大きい語をそれぞれの概念グループの代表語とする．

最も重要度が大きい語は，それぞれの概念グループで最もユーザに使用される重要な語だと判断できるため，概念グループの代表語としても問題ない．ここで，商品紹介の概念グループ $G_i^e(Q_i^j)$ の代表語を $w_f(G_i^e)(f = 1,2,3,\ldots)$，レビュー文書の概念グループ $G_i^l(R_i^k)$ の代表語を $w_h(G_i^l)(h = 1,2,3,\ldots)$ とすると，これらが商品に対する重要語になる．

商品紹介とレビュー文書の代表語 $w_f(G_i^e)$，$w_h(G_i^l)$ を決定する際に用いられる重みは，式 (3) で示した TF-IDF 値である．各文書において各概念に属する $w_q(Q_i^j)$，$w_r(R_i^k)$ が判明しているため，それぞれの TF-IDF 値を計算した上で最も大きな TF-IDF 値を持つ語が代表語として選出される．これにより，その商品に特有の代表語を決定することができている．

なお，TF-IDF 値を計算する際，商品紹介ページ Q_i^j とレビュー文書 R_i^k では，それぞれの文書を編集した人数が異なるため，重み計算に利用する代表ページを選出する必要がある．たとえば商品紹介ページ Q_i^j は，通常，単一ページで構成されるため，最も更新日時が新しい商品紹介ページ $Q_i^{j'}(j' \in j)$ を用いて TF-IDF 値の計算を行う．これに対して，レビュー文書 R_i^k は，複数の消費者の視点からの主張を含む文書であるので，すべての文書の TF-IDF 値を考慮すると，投稿数の多い消費者の主張が他者の主張よりも，強く反映される可能性がある．そこで，すべてのレビュー文書投稿者の主張を均等に反映するために，レビュー文書投稿者一人に対して最新日時の代表レビュー文書 $R_i^{k'}(k' \in k)$ を利用する．

3.5　タグクラウドの作成

3.4 節で説明したように，商品紹介ページおよびレビュー文書それぞれから抽出した代表語群を視覚的に表現するために，それぞれのページ，文書ごとに二種類のタグクラウドを作成する．タグクラウドの作成には，プログラミング

[*6] http://wordnet.princeton.edu/man/wninput.5WN.html#toc 2013 年 11 月閲覧
[*7] http://www.cs.cmu.edu/~hideki/software/jawjaw/ 2013 年 11 月閲覧

言語 JavaScript を用いて開発されている d3-cloud [*8] を用い，タグの大きさは代表語の TF-IDF 値の大きいものほど大きく表示されるように文献 [9] を参考に 10pt〜100pt の間で設定した．このように文字サイズを設定することにより，タグの大きさに変化が出つつ，小さすぎるタグも見ることがサイズとなる．

4. 評価実験

EC サイトにおいて，EC サイトユーザにとって商品情報の有益なリソースは各商品の紹介ページと消費者によるレビュー文書である．本稿で実装したシステムは 1 節でも述べたように，爆発的に増えた商品とそのレビュー文書を整理し，EC サイトユーザがリソースを情報として利用でき，最終的には購買行動の促進につなげるためのものである．このことから，購買行動促進システムの評価実験として行うべきことは，

(1) タグクラウドに表示される代表語の抽出を，ゼロ代名詞を考慮して行うべきであったかどうか
(2) タグクラウドの表示法として，商品紹介ページだけ，レビュー文書だけ，商品紹介およびレビュー文書をミックスして，商品紹介およびレビュー文書を別々に使用するという四種類のうち，どの手法が最適か

を調べることである．双方とも，主観的評価を行う必要があるため，代表語の抽出の是非やタグクラウドの表示法の可否は被験者が対象商品を購入する決断ができるか否かで判断することにした．

なお，評価実験に使用したデータセットは，楽天レビューデータセット [*9] を用いており，約 5,000 万件の商品データから母比率の区間推定から求めた最低必要サンプル数である 400 件をランダム抽出し，それらの商品紹介ページとレビュー文書を元に提案手法を用いてタグクラウドを作成した．

4.1 代表語抽出の是非

本実験においては，提案手法のようにゼロ代名詞を考慮して代表語の TF-IDF 値を計算した場合とそうでない場合を比較して，レビュー文書から生成されたタグクラウド，すなわちレビュー文書から抽出された代表語の質を評価した．

具体的には 102 名の被験者に対し，50 商品のタグクラウドと商品紹介ページを見てもらい，その商品を身内のために購入できるか否か，つまりその商品の購入に関する意志決定ができるか否かを調査した．被験者には同一商品のタグクラウドを閲覧することはなく，また，ゼロ代名詞を考慮しているかどうかもわからないよう工夫し，「購入する」「購入しない」「購入するか否か判断できない」の三段階の指標に基づいて質問紙調査を行ったところ，表 1 のような結果となった．この結果に対して自由度 2，有意水準 5% でカイ二乗検定を行ったところ，カイニ乗値が 162.54 となり，その値をとる確率は 1% 未満 ($p < 0.01$) となった．

表 1 代表語抽出に関する調査結果

	購入する/購入しない	判断できない
ゼロ代名詞未考慮	1393	1157
ゼロ代名詞考慮 (提案手法)	1832	718

このため，設定した帰無仮説「比較対象タグクラウドと購買判断が独立である」が棄却され，「比較対象のタグクラウドと購買判断に関連がある」ことが判明した．また，どちらの方法で代表語を抽出したほうが有用であったかを判断するために，表 1 の数値を比較したところ，提案手法を用いて作成したタグクラウドの方が商品の購買を判断できなかった人数が少なかったため，ゼロ代名詞を考慮して代表語を抽出した方が有用であることがわかった．

4.2 構築したタグクラウドの是非

本実験では，400 件の商品に関する商品紹介ページとレビュー文書を用いて，

- 商品紹介ページのみでタグクラウドを構築
- レビュー文書のみでタグクラウドを構築
- 商品紹介ページとレビュー文書から代表語を取り出し，それらを用いて一つのタグクラウドを構築
- 商品紹介ページとレビュー文書から代表語を取り出し，それぞれ別々のタグクラウドを構築 (提案手法)

を行うことで，タグクラウド構築の評価を行う．これは，本稿で提案したタグクラウド構築法の優位性として，短時間で販売側と消費者の意見を比較評価することができることが挙げられるためである．したがって，他三種のタグクラウド構築法とカイ二乗検定を用いて比較評価し，4.1 節と同様，どのタグクラウド構築法が商品を身内のために購入するか否かの意志決定の役に立つかを調査した．表 2 〜 4 にそれぞれの調査結果を示す．

表 2 タグクラウドの比較 (商品紹介ページのみ vs 提案手法)

	購入する/購入しない	判断できない
商品紹介ページのみ	1635	915
提案手法	1940	610

表 3 タグクラウドの比較 (レビュー文書のみ vs 提案手法)

	購入する or 購入しない	判断できない
レビュー文書のみ	1524	1026
提案手法	1940	610

これらの結果に対して自由度 2，有意水準 5% でカイ二

[*8] http://www.jasondavies.com/wordcloud/ 2013 年 11 月閲覧
[*9] http://www.nii.ac.jp/cscenter/idr/rakuten/rakuten.html 2013 年 12 月確認

表 4　タグクラウドの比較 (統合手法 vs 提案手法)

	購入する/購入しない	判断できない
統合して表示	1856	694
提案手法	1940	610

乗検定を行ったところ，カイ二乗値とその p 値は表 5 のようになった．このため，設定した帰無仮説「比較対象のタグクラウドと購買判断が独立である」が棄却され，「比較対象のタグクラウドと購買判断に関連がある」ことが判明した．また，どちらの方法でタグクラウドを表示した方が有用であったかを判断するために，各表内の数値を比較したところ，提案手法を用いて作成したタグクラウドの方が商品の購買を判断できなかった人数が少なかったため，提案手法に基づいてタグクラウドを構築した方が有用であることがわかった．

表 5　カイ二乗値と p 値

	カイ二乗値	p 値
表 2	87.02	$1.07 \times 10^{-20} < 0.01$
表 3	155.74	$9.66 \times 10^{-36} < 0.01$
表 4	7.01	$7.01 \times 10^{-3} < 0.01$

以上三つのカイニ乗検定の結果，四種類の表示方法の中で商品紹介ページとレビュー文書それぞれを別々にしてタグクラウドを構築する方法が最も EC サイトユーザにとって購買判断しやすい表示方法ということが判明した．

4.3 考察

以上の結果より，EC サイトの商品紹介ページやレビュー文書を用いてタグクラウドを構築する場合は，文書自体が持つ顕現性を利用し，ゼロ代名詞を考慮した代表語の抽出をすべきであることが判明した．また同時に，商品紹介ページやレビュー文書といった二種類の異なる文書からタグクラウドを構築するためには，それぞれ別のタグクラウドを構築することが EC サイトユーザにとっては有用であることも判明した．これは，本稿の提案が語の類似性を日本語 WordNet を用いることで同一概念をグルーピングし，表記の揺れを軽減させたことによる影響も大きい．

しかしながら，語の多義性については未考慮であったため，たとえば「ワンピース」という名詞がタグになる場合，指し示すページが「洋服」の意味でその語を使用しているのか，もしくは「漫画」のタイトルを表しているのか，「ワンピース」という表記だけからは判断できず，商品購買の意志決定につながらなかったケースがあった．そのため，抽出してきた代表語に関連のある語を併せて抽出し，その二語間の関係から代表語の持つ意味を詳細に把握する仕組みを構築することが今後必要となる．

5. おわりに

本稿では，係り受け解析器，日本語概念辞書，データ視覚化技術を用いて，EC サイトの商品紹介ページやレビュー文書から重要語抽出し，それらをタグクラウドを用いて提示する方法を提案した．また，提案方法の評価のために，構築したタグクラウドの提示が商品の購買行動につながるかどうかを確かめたところ，従来手法で用いられていた代表語の抽出手法やタグクラウドの構築法よりも，提案手法のほうが有用性があることがわかった．本稿の提案により，専門家がタギングを行っていた時のような表記の揺れの少ない代表語によるアノテーションと，複数人で行うソーシャルタギングのようなユーザの主張を反映したアノテーションという，双方のメリットを組み合わせたタグが機械的に抽出できる可能性がある．

しかし，提案手法を用いることでタグに対する表記の揺れはなくなったが，意味を複数持つ語がタグになった場合，語の意味を特定することができず，効果的なタグクラウドを構築することができないという問題が残されている．この問題を解決するために，抽出してきた代表語の中にある名詞とその名詞を修飾している形容詞の対から，その名詞が何を意味している語なのかを判断できる仕組みを構築することが今後必要となる．

参考文献

[1] Suchanek, F. M., Vojnovic, M. and Gunawardena, D.: Social Tags: Meaning and Suggestions, *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ACM, pp. 223–232 (2008).
[2] 笹野遼平，黒橋禎夫：大規模格フレームを用いた識別モデルに基づく日本語ゼロ照応解析，情報処理学会論文誌，Vol. 52, No. 12, pp. 3328–3337 (2011).
[3] Isahara, H., Bond, F., Uchimoto, K., Utiyama, M. and Kanzaki, K.: Development of the Japanese WordNet, *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, ELRA, pp. 2420–2423 (2008).
[4] Baeza-Yates, R. and Ribeiro-Neto, B.: *Modern Information Retrieval: The Concepts and Technology behind Search*, Addison-Wesley Professional, 2nd edition (2011).
[5] 飯田　龍，徳永健伸：談話の顕現性を考慮した重要語抽出とその応用，情報処理学会研究報告，Vol. 2009-NL-193, No. 9, IPSJ, pp. 1–8 (2009).
[6] 馬場雪乃：ソーシャルタギングからのことばが指し示す実世界対象の表現獲得，博士論文，東京大学情報理工学系研究科 (2012).
[7] Fujita, A., Isabelle, P. and Kuhn, R.: Enlarging Paraphrase Collections Through Generalization and Instantiation, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ACL, pp. 631–642 (2012).
[8] Kaser, O. and Lemire, D.: Tag-Cloud Drawing: Algorithms for Cloud Visualization, *Proceedings of Tagging and Metadata for Social Information Organization* (2007).
[9] 下村香理，芦澤昌子，佐川　賢：高齢者の文字可読性に及ぼす色および照度レベルの影響，日本色彩学会誌，Vol. 36, No. 1, pp. 15–26 (2012).