Web上の文書を用いた商品の評判記述システムと構築手法の提案

Web上の文書を用いた商品の評判記述システムと構築手法の提案

Web上の文書を用いた商品の評判記述システムと構築手法の提案

澤井政宏*，岡田吉史**，長島知正***
*室蘭工業大学 生産情報システム工学専攻(D3)，
産業技術総合研究所 生命情報科学研究センター，*室蘭工業大学 情報工学科

1. はじめに

今日、Internetを代表とする各種メディアの発達やオンラインショップの増加に伴い、商品に関する情報を手軽に入手することや、以前とは比べ物にならないほど多くの商品の中から、必要とする商品を選択することができるようになった。しかし、商品の選択肢が増加したがゆえに、実際に既存のシステムを利用して商品の情報を調べ、調べた情報をもとに自らの好みに合った商品を決定することは非常に手間のかかる作業となっている。このような背景の下、消費者の好みを把握する技術[1]や、把握した好みに基づいた商品検索システム、推奨システムの開発が盛んに行われている[2][3][4]。

我々は商品を購入する際、他者から商品の評判を聞いて、自らの好みに合う商品か否かを判断する場合がある。これは、商品が好みに合うかを判断する上で、他者の商品に対する評判が重要な情報の１つになっていることを示唆している。よって、商品の評判を計算機処理可能な形式で記述することができれば、商品検索や推奨システムの情報源として利用することができると考えられる。そこで本研究では、商品の例として音楽アーティストを取り上げ、Web上のアーティストに対する評判が記述された文（以下、評判文）を用いて、アーティストの評判を記述したデータベースを構築する手法を提案し、実装することによって、実装システムの評価を行った。

2. 評判文

本研究では、評判文の例として、Amazon.co.jpから収集した10名のアーティスト（表１）のCDに対する評判を記述した文を考える。10名のアーティストは全て、カテゴリ「J-POPアーティスト」に属するアーティストである。Amazon.co.jpでは、表２のように各文を５段階評価しているが、本研究では、好みを問題とするため、３以上の肯定的な評価点を持つ文（文数18485）を対象のデータとする（よって、表２における評価点が１の文は対象外のデータとする）。これらはアーティストのCDに対する評判を表す文ではあるが、CDに対する評判はそのCDを創作したアーティストに対する評判の一部と見なすことができるため、この文をアーティストに対する評判文として扱った。表２の評判文が示すように、これらの中には「歌詞」といったアーティストを好む要因となったアーティストが持つ属性と、「ポジティブな」といったその要因（属性）に対する評価が記述されている。

3. 評判を記述したデータベース

2章で示した評判文は、計算機にとっては単なる記号の羅列に過ぎないため、記述されたアーティストに対する評判を計算機可読な形式で表現する必要がある。一般に、商品の評判は「アーティストAの歌詞が素晴らしい」や「ポジティブなメロディー」といったように、対象物（アーティストA）と属性（歌詞、メロディー）、属性に対する評価（素晴らしい、ポジティブな）から構成される。そこで本研究ではこれら３つを用いて、木構造により商品の評判を記述する。

表 １：評判文を収集した10名のアーティスト

B’z、aiko、宇多田ヒカル、Mr.chirdren、Chemistry、浜崎あゆみ、平井堅、ゆず、175R、every little thing

表 ２：評判文の例

評判文	評価点
軽快なリズムの中に、B’z本来の音楽性が盛り込まれていて聴いていて気持ちがよくなる。今回の曲はB’z独特のハードロック調では・・・	4
松本さんのパワフルなサウンドに稲葉さんのポジティブな歌詞が重なり合ったホントにB’zらしい曲です。聞いているだけで元気になっ・・・	5
なんとなく「マニアック性」「固定客向け」的な音楽がイマイチのれません。工夫を凝らしたというよりは、マニアックに富んでいるという表・・・	1

図１に本研究で構築するデータベースを示す。データベースは、図１上のように、商品ごとの評判を格納したレコードを持つ。各商品の評判は、図１下のような評判木によって表現される。評判木は、単一の対象物が持つ属性の上下関係と、属性に対する評価を木構造により表したものである。四角で示したノードは属性を示し、楕円で示したノードは評価表現を表す。実線の矢印で示したエッジは属性間の上下関係を示し、破線の矢印で示したエッジは属性と評価表現間の修飾関係を示す。rootは対象とするアーティストを表す。嗜好木内の各属性は、その属性に対する評価を表すサブツリーを持つ。このようなサブツリーを持つ点が、嗜好木が単純な木構造と異なる点である。例えば、属性「曲」の下位属性として「メロディー」、「歌詞」、「リズム」が存在したとする。「曲」の評価は曲の構成要素である「メロディー」「テンポ」「リズム」などの評価から成り立っていると考えられる。本嗜好木は、上記した構造を持つことにより、上位の属性に対する評価が下位の属性の評価から成り立っていることを記述できる。

4. 提案方法

本章では３章で示したデータベースを半自動的に構築する方法を提案する。4.1 節では評判文における対象物、属性、

評価表現はどのようなものであるかを説明し、4.2.節～4.4.節ではデータベースを構築する具体的な方法を説明する。

4.1. 評判文における対象物、属性、評価表現

対象物とは、消費者（ここでは、評判文の記述者）が好む商品や人物のことであり、本研究で対象とする評判文においてはアーティストである。属性は、対象物を好む要因となった対象物が持つある一つの特徴であり、評判文においては、J-POPアーティストというカテゴリに属するアーティストを好む要因を与える名詞単語であるものとする。評価表現は、属性に対する評価であり、評判文においては属性を直接修飾する形容詞単語とする。例えば、アーティストAに対する評判文中に「声が素晴らしい」という文があった場合、対象物はアーティストA、属性は「声」、評価表現は「素晴らしい」となる。

4.2. 対象物を好む要因となっている属性の特定

2章で示した全ての評判文（10名のアーティストに対する評判文全て）中に現れる名詞単語の出現頻度と共起関係に基づいて、属性の特定を行う。まず初めに、10名のアーティストに対する評判文に対してchasen[5]を用いて形態素解析を行い、文を単語単位に分割し、品詞情報を付加した。アーティストに対する評判文の中で、高い出現頻度を持つ名詞単語は、それらのアーティストの属性とその属性に対する評価を語る上でよく用いられる名詞単語であるため、属性を表す単語である可能性が高い。また収集した評判文の中には、アーティストに対する評価以外の話題を記述した文も含まれているが、同一文中に解析者に属性と見なされた単語（以下、選択属性）が多く含まれるならば、その文はアーティストの属性とその評価について記述されている可能性が高い。すなわち、一文の中で、選択属性と頻繁に共起する名詞単語は、属性となる可能性が高いと考えられる。そこで、名詞単語tが属性として尤もらしい度合いを表す属性度F(t)を、tの出現頻度TF(t)とtが選択属性と共起する度合いA(t)を用い、次式より算出する。

$$F(t) = \mathrm{RANGESCALE}(\log TF(t)) * \mathrm{RANGESCALE}(A(t))$$

$$A(t) = t\text{を含む文中に存在する選択属性の数}/TF(t)$$

ここで、RANGESCALEは補正したTF(t)とA(t)を最大値１、最小値０に正規化する処理を表す関数である。評判文に表れる単語の出現頻度は、少数の単語が大きな値を持ち、それ以外の単語は比較的小さな値を持つ傾向にある。そのため、TF(t)の値を直接属性度に反映した場合、少数のTF(t)の値が高い単語が大きな属性度を持つことになる。そこで、logを用いてTF(t)を補正した。これらの値を正規化し、掛け合わせることによって属性度F(t)を算出する。

本手法では、以上のように算出された属性度に基づいて名詞単語をランキングし、その結果から属性としてふさわしい単語を解析者が手作業により選出する。一般に、属性はアーティストごとに異なると考えられる（例えばアーティストAはギターを演奏するため属性「ギター」を持つが、アーティストBはギターを演奏しないため属性「ギター」を持たないなど）。しかしながら、ある属性を対象物が持たない場合、その属性に対する評価は評判文中には出現しないので、その属性に対する評価表現は結果的に0になる。従って、対象物が持つ属性の差異は、属性に対する評価表現の数で区別できる。よって、属性は対象とするカテゴリ（本研究においては音楽アーティスト）に属する対象物が持つ属性を網羅するように選出することが望ましい。

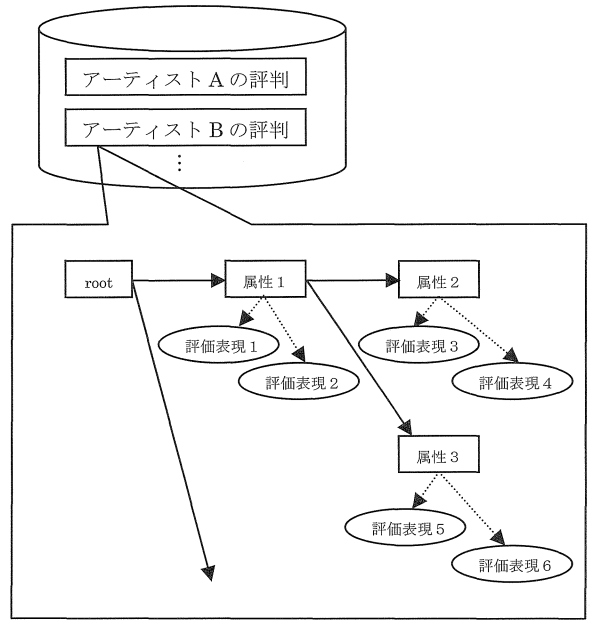

図１：構築するデータベース

4.2. 属性の上下関係の推定

4.2.節で得られた属性の上下関係を推定する。我々は相関ルールマイニング[6]を用いて、属性の上下関係の推定を行った。相関ルールマイニングとは、共起するアイテム同士のパターンを表す相関ルールを導出する手法である。アイテム集合を$I=\{i_1, i_2, \cdots, i_m\}$、トランザクションデータベースをDとする。各トランザクションＴはＩの部分集合である。導出される相関ルールは$X \Rightarrow Y (X, Y \subseteq I, X \cap Y = \Phi)$で表現される。D全体に対してXとYを共に含むトランザクションの割合を支持度、Xを含むトランザクションの内XとYを共に含むトランザクションの割合を確信度と呼び、これら２つの値によってルールの重要度が見積もられる。我々はアイテムを属性、トランザクションを１文に出現する属性の集合として相関ルールマイニングを行った。例として、曲⇒声（支持度20%、確信度70%）という相関ルールは「曲と声を共に含む文は全文中の20%であり、曲を含んでいる文の70%が声を含む」ことを意味する。相関ルール導出のためのアルゴリズムはいくつか提案されているが、我々はAgrawalらによって提案されたaprioriアルゴリズム[7]を用いた。

相関ルールX⇒Yが存在するとき、同じ支持度を持ったY⇒Xという相関ルールが存在するが、これらは通常異なった確信度を持つ。相関ルールX⇒Yの確信度に対して、相関ルールY⇒Xの確信度が十分大きい場合、「Yが出現するときはXを伴う傾向が強いが、Xが出現するときはYのみならず他の多くの属

表 3：アンケートにより得られたアーティストを好む要因

声、メロディー、テンポ、リズム、歌詞、楽器

表 4：属性として選出した 11 の単語

曲、歌詞、声、ギター、メロディー、歌、テンポ、ピアノ、リズム、ソロ、ドラム

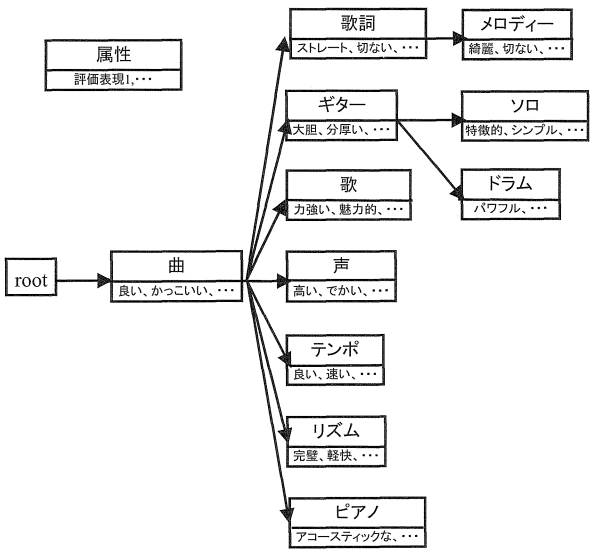

図 2：提案手法で構築した評判木（一部）

性も伴う」ことを意味する。このため、YはXの下位属性である可能性が高いと考えられる。そこで我々は相関ルールX⇒Yの確信度に対して相関ルールY⇒Xの確信度が1.5倍以上である場合、YはXの下位属性とした。ここで、問題の単純化のため、各属性は単一の上位属性を持つという制限を設けた。上記の方法により上位となる属性が複数存在する場合は、確信度が最も高い相関ルールを用いて上位属性を判断する。例えば、属性Yの上位属性がA、B、Cの３つ存在し、相関ルールY⇒A、Y⇒B、Y⇒Cの確信度がそれぞれ20%、60%、10%であった場合、属性Yの上位属性はBとなる。もし、上位属性が存在しなかった場合には、対象物を表すrootの下位属性とした。

4.4. 評価表現の収集

4.3.節で得られた属性の上下関係を表す木に評価表現を付加し、評判木を生成する。我々はアーティストごとの評判文に対して日本語係り受け解析器cabocha[8]を用いて係り受け解析を行い、属性と係り受け関係にある単語をアーティストごとに抽出した。評価表現は属性を直接修飾する形容詞であるため、属性と直接係り受け関係にある単語のうちcabochaによって付加された品詞タグが「形容詞－自立」、「名詞－形容動詞語幹」である単語を抽出した。また、名詞に続いて出現する場合、形容詞的な働きをする単語である「形容詞－非自立」、「形容詞－接尾」、「名詞－接尾－形容動詞語幹」の品詞タグを持つ単語も名詞とあわせて抽出した。このような評価表現を、4.3.節で得られた木の、係り受け関係にある属性に結び付けることによって、アーティストごとの評判木を生成する。

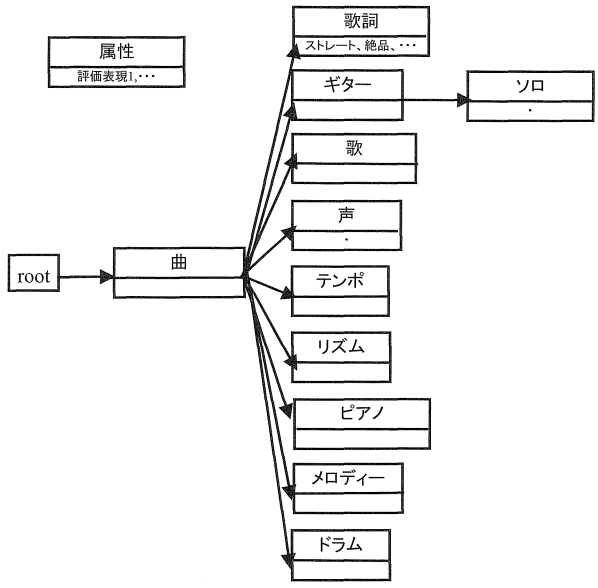

図 3：人手で構築した評判木（一部）

5. 提案手法の評価

本章では構築した評判木が、記述者の商品に対する評判を表現できているかを評価することにより、提案手法の評価を行う。評価は、提案手法によって構築された評判木と、評判文を読んで人手によって構築した評判木を比較することによって行う。

5.1. 評価方法

まず初めに我々は、20 名の被験者に対して、自由記述形式で好きなアーティストとそのアーティストを好む理由を尋ねるアンケートを行った。我々は得られた回答を人手により整理し、表３のような６つのアーティストを好む要因を得た。次に、4.2.節に示した方法を用いて、上記の６つの要因に該当すると思われる 11 の単語（表４）を属性として選出した。そして 4.3.～4.4.節に示した方法でアーティスト B’z に対する評判木を構築した。構築結果の一部を図２に示す。

次に我々は、評判文を被験者に読ませ、上記の 11 の属性の上下関係を表す木を構築させた。さらに、属性「歌詞」に対する評価を表していると考えられる単語を収集させ、人手による評判木を構築した。人手によって構築された評判木を図３に示す。

評価は、1)提案手法で構築した評判木が、人手による属性間の上下関係をどの程度再現しているか、2)提案手法で収集した属性「歌詞」に対する評価表現（以下、評価表現・提案）が、人手によって収集された属性「歌詞」に対する評価表現（以下、評価表現・人手）とどの程度一致するか、の2点について行う。1)については、人手による上下関係を正解としたときに、提案手法にどの程度誤った上下関係が含まれているかを調べることにより評価を行う。2)については、再現率、適合率により評価を行う。本評価における再現率、適合率は以下の式で算出される。

$$再現率=\frac{R}{C}$$
$$適合率=\frac{R}{N}$$

R：評価表現・人手と一致する評価表現・提案の数
C：評価表現・人手の数
N：評価表現・提案の数

5.2. 評価結果と考察

5.2.1. 属性の上下関係の比較結果と考察

人手による属性の上下関係（図３）を正解とした場合、提案手法の上下関係の推定（図２）では、「歌」、「声」、「テンポ」など９個の属性は「曲」の下位属性として正しく推定されたが、残る「メロディー」、「ドラム」については、「メロディー」が「歌詞」の下位に、「ドラム」が「ギター」の下位属性として推定され、推定は失敗している。本システムでは属性間の上下関係を相関ルールの確信度の差に基づいて推定している。この方法では、ある属性とその下位となるべき属性を同時に含んでいる評判文（例文：「曲」の「歌詞」が良い、この「曲」は「テンポ」がここちよい）が多い場合は適切な推定結果を得ることが出来る。しかし、本来上下関係のない属性を同時に含んでいる評判文（例文：「歌詞」がよく「テンポ」もよい、「歌詞」と「メロディー」が切ない）が多数存在する場合には、それらがノイズとなり、上下関係の推定を誤りやすい。以上より、相関ルールの確信度によって属性間の上下関係を推定する本手法は、ある程度正しい推定を行うことが出来るが、同時にそれだけでは十分ではないことも示している。

5.2.2. 評価表現の比較結果と考察

表５に再現率、適合率を示す。適合率は83%と比較的高い値であるが、再現率は55%と低い値を示している。評価表現・提案として収集されたが、評価表現・人手に含まれなかった例として最も多かったものは、「ダンスビートにシリアスな歌詞をいれ、B'z伝説の始まりとなった作品」といった文の中の「シリアスな」といった形容詞である。このような文は評判文記述者のアーティストに対する評価が記述された文ではなく、アーティストの背景情報を記述した文である。しかし、本システムでは、このような文からも評価表現を収集してしまう。適合率をさらに向上させるために、評判文記述者自信の評価を表している文を同定することが必要と考えられる。また、評価表現・提案として収集されなかったが、評価表現・人手に含まれていた例として、「相変わらず稲葉さんの歌詞は絶品です」、「切ないメロディーと奥の深い歌詞が涙を誘います。」といった文の「絶品」、「涙を誘う」などのような、名詞や目的語を伴った動詞によって構成される評価が挙げられる。本手法で収集できなかった評価表現・人手のほとんどはこのケースであった。本研究では評価表現を、属性を直接修飾する形容詞としたが、名詞や目的語を伴った動詞が評価を表すこともある。そのため、形容詞以外の品詞を持つ評価も抽出する必要があると考えられる。

表 7：提案手法で収集した属性「歌詞」に対する評価表現の再現率・適合率

再現率	適合率
54.54545455	83.07692308

7. おわりに

本論文では、Web上の評判文を用いて、評価者の嗜好を対象物の属性に対する評価によって計算機処理可能な形で記述する手法を提案した。評価実験により、本手法は１）属性の上下関係を推定できること、２）評判文に記述された属性に対する評価を表現可能なことを示した。しかしながら、提案手法の評判木構築精度は十分なものではなく、更なる精度向上が必要であることも明らかになった。

本論文では対象物としてアーティストを取り上げたが、アーティストに限らず家具や車など多くの対象物は、対象物の持つ属性が消費者の嗜好に強く影響を与えていると考えられる。このような対象物に対する嗜好を記述することは、嗜好という感性情報を取り入れたシステムを開発する一つの切り口になると考えられる。

参考文献

1) 村上知子,酢山明弘,折原良平：ベイジアンネットワークを用いた消費者行動モデルの構築実験,第18回人工知能学会大会, 3F3-01, 1-4, 2004
2) Resnick P., Iacovou N. Suchak M. Bergstorm P. and Riedl J. GroupLends : An Open Architecture for Collaborative Filltering of Netnews,Proc. ACM Conf. On computer Supported Cooperative Work, 175-186, 1994
3) 矢野絵美,北野有亮,末吉恵美,篠原勲,ピンヤポンシニーナット,加藤俊一：消費者の感性モデルを利用したレコメンデーションシステムの構築,情報処理学会論文誌, 44, SIG8, 46-54, 2002
4) 池添剛,梶川嘉延,野村康雄：音楽感性空間を用いた感性語による音楽データベース検索システム,情報処理学会論文誌, 42, 12, 3201-3212, 2001
5) 松本祐治：日本語形態素解析器chasen, http://chasen.naist.jp/hiki/ChaSen/
6) Agrawal R., Imielinsi T. and Swami A.: Mining Association Rules between Sets of Items in Very Large Databases, Proc.ACM, SIGMOD, Int. Conf. Management of Data, 207-216, 1993
7) Agrawal R. and Srikant R.: Fast Algorithms for Mining Association Rules in Large Databases, Proc. 20th Int. Conf. Very Large Data Bases, 478-499, 1994
8) 奈良先端科学技術大学自然言語処理学講座：日本語係り受け解析器cabocha, http://chasen.org/~taku/software/cabocha/