比較評価情報に基づくランキング手法

Ranking Method using Comparative Relations

倉島 健♥ 別所 克人♥ 戸田 浩之♥
内山 俊郎♥ 片岡 良治♥ 奥 雅博♥

Takeshi KURASHIMA Katsuji BESSHO
Hiroyuki TODA Toshio UCHIYAMA
Ryoji KATAOKA Masahiro OKU

比較評価文は対象間の優劣を述べた表現である．対象（典型的には商品）を体験した消費者が発信する大量のテキストデータからこれらの情報を収集し分析することができれば，本当に価値のある対象を効果的に発見することができるはずである．本稿では，Consumer Generated Media(CGM) から人々が対象間の優劣を述べた比較評価情報を抽出し，それをもとに対象をランキングする手法を提案する．対象をグラフのノードで，比較評価情報の抽出結果集合から導き出された対象間の優劣関係をグラフの有向辺で表現したグラフを生成し，その構造を解析することで各々の対象の評価値を算出する．このグラフは「より良い商品を求めて売り場を移動する顧客」の行動モデルに基づいて生成する．

A comparative sentence expresses a relation between two objects. By mining the information from a large number of documents generated by consumers, we can rate the values of objects efficiently. This paper proposes a method of mining the comparative relations and ranking the objects based on mined results. We generate the graph, each node corresponds to a object, and each edge expresses the relation between two objects. By applying the PageRank method to the graph, the system brings order to the object. This type of graph can be thought of as modeling the behavior of a potential customer.

1. はじめに

近年，人々は商品を実際に体験した消費者の反応を参考にして商品を選択するようになってきている．しかし，一般に消費者が発信するそのような情報は膨大であるため，大量のテキスト情報から如何にして人々の反応，評価を抽出し，そして集約するかが評判分析分野の重要な課題となっている．

♥ 正会員 日本電信電話株式会社 NTT サイバーソリューション研究所 {kurashima.takeshi, bessho.katsuji, uchiyama.toshio, kataoka.ryoji}@lab.ntt.co.jp

従来，評判分析分野においては「映画AのBGMは良い」といった単一の対象に対する評価を，対象，属性，評価という３つ組の意見モデルを定義して抽出してきた[1][2]．このような単一の対象に対する単体評価は評価者の経験に左右されやすいという欠点がある．その結果，不特定多数の評価者の単体評価を集約したランキング結果も信頼性が高いとはいえなかった．例えば，甘く評価をつける人が「良い」という評価をした対象Xと，厳しく評価をつける人が「悪い」という評価をした対象Yとで一概に「Xの方がYよりも良い」と結論付けることはできない．一方で，個人が発信する文書の中には「映画AのBGMは映画Bよりも良い」といった複数の対象の優劣を述べた比較評価文が存在する．比較評価文は２つの対象の相対的な位置関係のみを表現する．AとBの両者を実際に体験した評価者によって述べられたこの情報は，両者の優劣関係を正しく把握するための重要な鍵であるといえる．消費者が発信した大量のテキストデータからこれらの情報を収集しまとめあげることができれば，本当に価値のある対象（商品）を効果的に発見でき，その結果も他者の同意を得やすいものになるはずである．

本研究においては，人びとが複数の対象を比べあわせて，そこに認められる優劣を述べた比較評価情報を抽出する．比較評価情報は｛評価対象，比較対象，属性，評価｝から構成され，本研究においては，このうちの｛評価対象，比較対象，評価｝に焦点を絞って抽出を行う．属性の抽出については，従来の３つ組の評判抽出における主要課題であり，今後はこれらの研究を参考にして抽出する予定である．なお，今回は特に映画ジャンルに絞って比較評価情報の抽出実験を行う．本研究では，さらに，比較評価情報抽出によって得られた２つの対象間の関係性から対象空間全体におけるそれぞれの対象の価値を算出する．対象をグラフのノードで，比較評価情報の抽出結果から導き出された対象間の関係をグラフの有向辺で表現したグラフを生成し，その構造を解析することで各々の対象の評価値を算出する．このグラフは「より良い商品を求めて売り場を移動する顧客」の行動モデルに基づいて生成する．得られた結果は，商品の購入を検討している潜在的な顧客や，マーケットアナリスト，あるいは商品を提供する側にとって有益な情報である．以降，２章で比較評価情報の抽出手法について，３章でランキング手法について，４章で評価実験の結果について，５章でまとめと今後の課題について述べる．

2. 比較評価情報の抽出

2.1 比較評価情報の定義

比較文とは，複数の対象の相対的な関係を述べた文である．Nitinら[3]の定義によると，英語における比較文は以下の４つのタイプに分類できる．

- *Non-Equal Gradable*:ある属性に関して複数の対象を優劣付けた表現（例）*greater, less than, -er -est.*
- *Equative*:ある属性に関して２つの対象が等しいことを述べた表現（例）*equal to, as as.*
- *Superlative*:複数の対象の中である対象が最も優れていることを述べた表現（例）*Greater than all others.*
- *Non-Gradable*:ある属性に関して複数の対象を比較した表現だが，明示的に優劣付けていない表現（例）*Toyota has GPS, but Nissan does not have.*

本研究においては *Non-Equal Gradable* に着目して抽出を行う．この表現に対応する日本語の表現を比較評価文，比較

評価文を構成する{評価対象，比較対象，属性，評価}を比較評価情報と呼ぶ．比較されている二つの対象の中で評価表現に意味的に主格で係っている方を「評価対象」，係っていないほうを「比較対象」とする．「属性」は，評価対象と比較対象を比較する際の評価項目であり，両者に共通する性質や特徴を示す表現である．「評価」は２つの対象間の優劣を述べた表現である．また，「評価」が肯定か否定のどちらの意味を持つかを示す「極性」は，テキストから抽出する情報ではないため,ここでは比較評価情報に含めないこととした．本論文においては，比較評価情報の中でも特に{評価対象，比較対象，評価}の抽出に焦点を絞って抽出を行う．

2.2 比較評価情報抽出の課題

本研究の目的は，映画や商品といったユーザが興味を持っている対象が，それが属するジャンル，概念階層においてどういった位置づけにあるかを導き出すことである．例えば，ある映画が，過去に上映された映画集合の中でどういった位置づけにあるかを知りたい．そのためには，ユーザが興味を持つ対象集合（初期対象集合）を入力として，それと同じ概念階層に属する対象間のすべての比較評価情報を抽出することが必要である．

本手法の基本的なアプローチは表層的なパターンを用いて比較評価情報を構成するそれぞれの要素を抽出することである．しかし，このような手法では，「評価対象」と「比較対象」の抽出において以下のような問題がある．

- ユーザが興味を持つ対象と異なる概念階層の比較評価情報までもが抽出されてしまう．例えば，映画について書かれたテキストには映画間の比較もあれば,俳優間の比較も存在する.ユーザが興味を持っている対象が映画であれば,映画に関する比較評価情報のみが必要である．
- 日本語の比較表現は,典型的には評価対象が「ほうが」，比較対象が「より」や「と比べ」などの表現で示されるが，比較文に特有でないパターンで「評価対象」と「比較対象」が出現した場合に抽出精度が低くなる.例えば，以下に示す例では「評価対象」である「映画A」と「属性」である「BGM」が同様の「は」格で出現する．
  （例１） 映画Aは映画Bよりも良い
  （例２） BGMは映画Bよりも良い

そこで，本手法においては，パターンマッチングで得られた候補語とユーザから与えられた初期対象集合とを照合することで高精度に「評価対象」と「比較対象」を抽出する．さらに，初期対象集合との照合のみでは，ユーザが与えた初期対象集合に含まれる対象しか「評価対象」「比較対象」として抽出できないため，照合を行う前に，初期対象集合をそれと同じ概念階層に属する対象集合へと拡張する．一般に，ユーザは，興味を持っている対象がどういった対象と比較されているのかを知らない，もしくは，知っていてもそれらが「前作」「旧作」等，正式名ではない表記でテキスト中に出現するといった場合がほとんどである．この拡張によりユーザが興味を持つ対象と同じ概念階層に属する対象間のすべての比較評価情報を抽出することが可能となる．初期対象集合の拡張は比較評価文の以下の性質を利用する．

- 「AよりBのほうが良い」というような評価対象，比較対象がともに比較評価文に特有な表現で出現した場合,AとBは高い確率で同じ概念階層に属する語である．

パターンマッチングの結果「評価対象」と「比較対象」が比較評価文に特有な表現で出現し，かつ，どちらか一方の語がすでに対象集合に含まれる語であった場合に，もう一方の語も対象集合に追加する．拡張した対象集合には，初期対象集合に含まれない対象や，対象が「前作」「旧作」等，正式名ではない表記で示される語が含まれることになる．

2.3 抽出アルゴリズム

(1)ユーザは入力として初期対象集合（ルート集合）$S=\{s_1,\ldots,s_m\}$を与える．典型的には，ルート集合はユーザがその価値を知りたいと考える複数の対象である．

(2)システムはルート集合に関連するWeb文書集合を収集する．文書を形態素解析し，解析データ$P=\{p_1,\ldots,p_n\}$を生成する．p_i=[単語$_1$/形態素$_1$]…[単語$_o$/形態素$_o$]である．

(3)p_iからパターンマッチングにより「評価対象」「比較対象」「評価」の候補語を取得し，以下の形式の初期トランザクションを生成する．

$T=\{x_1,..,x_m,y_1,..,y_n,t_{x1},..,t_{xm},t_{y1},..,t_{yn},z\}$
x: 評価対象の候補語 y: 比較対象の候補語
z: 評価の候補語 t: クラス値

クラス値は比較評価表現に特有のパターンで抽出された語とそうではない語を判別するためのもので，ここでは特有の場合に１を，特有でない場合には０をとる．クラス値は(4)の初期対象集合の拡張において必要となる．

(4)ルート集合Sからベース集合S'へと初期対象集合を拡張する．Tの各アイテムを順々に読み込んでいき，x，yがともに比較評価文に特有な表現で抽出された語(t_x=t_y=１)で，かつどちらか一方の語が，すでに対象集合に存在する場合，もう一方の語を対象集合に追加する．

(5)トランザクションTの各アイテムの評価対象x_i，及び比較対象y_iとベース集合S'とを，評価zと評価表現辞書とを照合し，成功した語を抽出する．評価表現辞書に格納されているデータの形式は{単語，形態素情報，極性}であり，照合を行うと同時に，極性の値を取得する．極性はP(肯定)，N(否定)のどちらかの値をとる．最終的に，システムはT'=(x',y',評価,極性)を出力する．

3. 対象のランキング手法

本章では，局所的な二項関係から，対象空間全体におけるそれぞれの対象の価値を導き出す手法を述べる．本手法においては「多くの良質な対象と比較されて相対的に評価を得ている対象」を評価の高い対象とする．良質な対象と比べて相対的に評価されている対象は，やはり良質であり，その数が多ければ多いほどその対象の価値は上がる.「相対的な評価」とは，２対象間の優劣の度合い・程度を表し，これは前章までに得られた比較評価情報の集合から導出する．そして，１つのノードが１つの対象を，ノード間の有向辺の重みに２つの対象の相対的な評価を反映させたグラフを構築し，生成したグラフの構造を解析することでそれぞれのノード（対象）の評価値を求める．

3.1 “顧客の行動モデル”に基づくグラフの生成

対象空間全体におけるそれぞれの対象の価値を導き出すために，前章で得られた二項間の優劣を参照しながら「より良い商品を求めて売り場を移動する顧客」の行動をモデル化する．顧客は商品群の中からある商品を見始め，次々に他の商品も見ていく．顧客の次の選択は大きく分けて(1) 現在見ている商品 A の前に留まり商品を見続ける (2) 他の商品 B の前に移動する である．この選択は，今見ている商品と他の商品との優劣に大きく依存する.現在見ている商品 A よりも優れている商品が少ない場合には，顧客はその場所にとどまって今見ている商品を見続ける．逆にいうと，他に優れて

いる商品が多く存在している場合には，他の商品の前に移動しやすい．次に，(2) の「他の商品 B の前に移動する」場合に，顧客がどの商品を選択するかについて述べる．顧客は現在見ている商品 A よりも相対的に支持されている商品ほど次に見る商品として選択しやすい．つまり，B が A より優れていると述べた人が，A が B より優れていると述べた人よりも多い B ほど，顧客は次に見る商品として選択する．

このモデルを表現するグラフを生成する．1 つのノードは 1 つの対象を表現し，ノード間の有向辺は対象間の関係性を表現する．有向辺は重み付きであり，あるノードから他ノードへの遷移確率は，それら対象（ノード）間の相対的な評価・優劣によって決まる．このグラフの遷移確率行列 A を以下に示す．今，ノード V の総数を N とする．A_{ij} はノード V_i からノード V_j に対する遷移確率であり，i と j は 1 から N の値をとる．

$$A_{ij} = \alpha S_{ji} / (S_{ji}+S_{ij}) \quad \text{if } (i \neq j)$$
$$= \beta W(i) / G(i) \quad \text{if } (i = j)$$
$$= 0 \quad \text{if } (i = j \text{ かつ } S_{ij} = S_{ji} = 0)$$

S_{ij}：ノード V_j に対するノード V_i の支持数
W(i)：ノード V_i の勝利数
G(i)：ノード V_i の対戦数
α：他ノードへの遷移確率に対する重み
β：自己遷移確率に対する重み

数式内で使われている用語の意味は以下の通りである．

支持数：対象Ａと対象Ｂ 間におけるＡの支持数とは B よりも A のほうが良いと述べた要素の数である．

対戦数：対象Ａの対戦数とは，対象Ａと比較された対象の総数である．

勝利数：対象Ａと対象Ｂ間において，Ａの支持数がＢの支持数よりも多かった場合に，対象Ａは対象Ｂに勝利すると呼ぶ．対象Ａの勝利数とは，対象 A が他の商品に勝利した回数である．

支持数 S_{ij}($i \neq j$)の値は前章までに得られたトランザクション T = (評価対象, 比較対象, 評価, 極性)の値から算出する．まずデータベースへの問い合わせによって対象集合 $O=\{o_1,o_2,...,o_N\}$を取得し，対角要素が 0 である N 次元正方行列 S を作成する．そして，S_{ij} を以下の計算式で算出する．

$$S_{ij} = \sup(O_i,C_j \Rightarrow P)+\sup(O_j,C_i \Rightarrow N)$$

$\sup(X \Rightarrow Y)$ はデータベース D 中においてアイテム集合 X と Y をともに含むトランザクション数である．O_i は対象 i が評価対象要素として出現した場合を，C_i は対象 i が比較対象要素として出現した場合を示す．この計算式で作成した行列 S の i 行目の要素の和は対象 i が優れていると述べられた総数であり，i 列目の要素の和は対象 i が劣っていると述べられた総数となる．W(i)と G(i)は，それぞれ S から求める．

$$W(i) = w_{i1} + w_{i2} + \cdots + w_{ij} + \cdots + w_{iN} \quad (j \neq i)$$
$$G(i) = g_{i1} + g_{i2} + \cdots + g_{ij} + \cdots + g_{iN} \quad (j \neq i)$$

w_{ij} は $S_{ij}>S_{ji}$ の時に 1 を，それ以外の時に 0 をとり，g_{ij} は $S_{ij}=S_{ji}=0$ のときに 0 を，それ以外のときに 1 をとる．最後に，行列 A の各要素を，その要素を含む行の和で割り，正規化する．そうして得られた A がグラフを表現する遷移確率行列であり，A_{ij} はノード V_i からノード V_j への遷移確率をあらわす．ノード V_i に対するノード V_j の支持数 S_{ji} が，ノード V_j に対するノード V_i の支持数 S_{ij} よりも相対的に多い場合に，$A_{ij} > A_{ji}$ となる．

3.2 評価値の算出

生成したグラフ（遷移確率行列）から，それぞれのノードの評価値 S(V)を求める．本手法においてはこの計算に PageRank を用いた[4]．PageRank の評価式を以下に示す．

$$S(V_j)=(1-d)\times\Sigma(A_{ij}\times S(V_i))+d$$

ノード V_j の評価値 $S(V_j)$はそれにリンクを張っているノードの評価値から求める．本手法で生成したグラフは重み付きグラフのため，ノード V_i からノード V_j へと伝播する値は，$S(V_i)$ に A_{ij} を乗じた値である．d は，ユーザが現在見ている商品（ノード）からまったく無関係な商品（ノード）の前に移動（ランダムジャンプ）してしまう確率である．PageRank をより直感的に述べると，ユーザがランダムに Web ページのリンクをたどり続ける行動のモデル（"random surfer"モデル）である．Web ページの評価値はランダムに Web 空間を歩き回り，最終的にそのページにたどりつく確率に等しい．つまり，提案手法によってある商品（ノード）に関して算出した評価値はより良い商品（ノード）を求めて売り場をまわる顧客が最終的にその商品（ノード）にたどり着く確率に等しくなる．

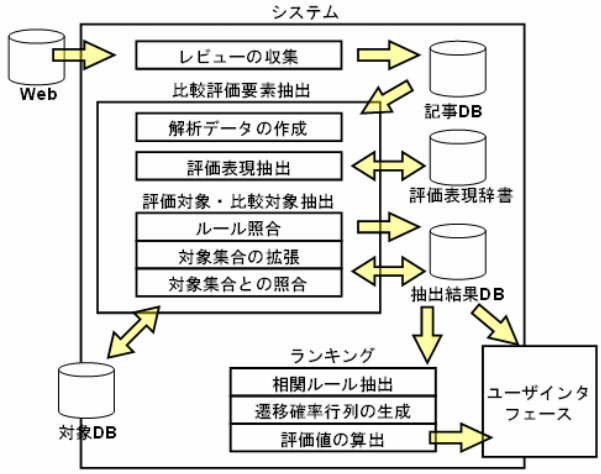

図１　システム構成
Fig 1: System configuration

4. 評価

4.1 比較評価情報の抽出精度

生成したシステムの構成を図 1 に示す．形態素解析には Jtag[5]を使用した．このプロトタイプシステムを用いて比較評価情報の抽出を行い，評価対象と比較対象の抽出精度を評価した．具体的には，以下に示す評価対象・比較対象抽出プロセスの組み合わせを考え，適合率，再現率，F 値を算出してどの程度精度が改善されるかを観察した．

(1) パターンマッチング
(2) 映画辞書との照合
(3) 拡張した対象集合 S'との照合

映画辞書には 1995 年 1 月～2006 年 12 月の日本で上映された映画の正式名称が格納されている．テキストデータはユーザレビュー500 記事（5 タイトル，各 100 記事)であり，初期対象集合 S は記事に関する映画 5 タイトルである．記事から人手で正解を作成した結果，63 記事に比較評価情報が含まれており，評価対象の要素数は 40，比較対象の要素数は 76 であった．表 2 に抽出結果を示す．評価対象，比較対象ともに，パターンマッチングに加えて映画辞書と対象集合との照合を行った(1)+(2)+(3)の F 値が最も高くなった．これは，辞書や対象集合との照合を行うことでパターンのみの抽出に比べて適合率が大きく改善したことに起因するもので

ある．また，(1)+(2)の映画辞書との照合においては再現率が0.1 程度と思ったほど効果が出ていない一方，対象集合との照合において再現率は 0.5 程度である．これは映画が正式名称で引用されることが少なく，省略された形，「前作」などの語で引用される頻度が高かったためである．比較対象よりも評価対象のほうがパターンマッチングのみの場合に比べて精度改善が大きい．これは，評価対象抽出においては，評価対象と属性がまったく同じパターンで出現することが多く，そういったケースで正解を抽出できているからである．

4.2 映画のランキング

ブログから抽出した比較評価情報の集合に基づき映画のランキングを行った．用いたデータは以下の通りである．

- 初期対象集合 S:期間中に上映中の映画 135 タイトル
- 拡張した対象集合 S':映画 462 タイトル
- 期間:2006 年 5 月 01 日～2007 年 1 月 31 日

ブログは，goo ブログ検索を用いて収集した．比較評価情報抽出の結果，「評価対象」，あるいは「比較対象」要素が欠けた場合には，ブログのタイトルに含まれる映画名を補足した．また，前章の評価において最も精度が高かった対象集合と映画辞書との照合を組み合わせた手法を用いた．対象集合の拡張によって得られた「１」「２」「前作」などの表現は，現時点では人手で映画名に変換している．抽出した比較評価情報集合をもとにランキングした結果を表３に示す．

表１　評価対象・比較対象の抽出結果
Table 1: Precision, recall and F-measure of extracted results

		(1)	(1)+(2)	(1)+(3)	(1)+(2)+(3)
評価対象	適合率	0.377	0.333	0.833	0.714
	再現率	0.725	0.100	0.500	0.625
	F 値	0.496	0.154	0.625	0.667
比較対象	適合率	0.737	0.889	1.000	0.981
	再現率	0.737	0.105	0.592	0.684
	F 値	0.737	0.188	0.744	0.806

表２　映画のランキング結果
Table 2: The results of movie ranking

映画名	評価値
嫌われ松子の一生	0.0170
DEATH NOTE デスノート　前編	0.0141
ゲド戦記	0.0133
LIMIT OF LOVE 海猿	0.0132
武士の一分（いちぶん）	0.0121
パイレーツ・オブ・カリビアン/デッドマンズ・チェスト	0.0120
硫黄島からの手紙	0.0119
カーズ	0.0100
ワイルド・スピード X3 TOKYO DRIFT	0.0100
どろろ	0.0094

5. まとめと今後の課題

本稿では，ＣＧＭから比較評価情報を抽出する手法と，その抽出結果をもとに対象をランキングする手法を示した．実験の結果，パターンマッチングと，初期対象集合の拡張手法によって得られた対象集合との照合を組み合わせることにより，比較評価情報抽出によって得られる「評価対象」「比較対象」の抽出精度が向上することを確認した．今後は，提案するランキング手法の特性を検証するとともに，属性（評価項目）の抽出も視野に入れてランキングを行う予定である．

[文献]

[1] 立石健二，福島俊一，小林のぞみ，高橋哲朗，藤田篤，乾健太郎，松本裕治. "Web 文書集合からの意見情報抽出と着眼点に基づく要約生成"，情報処理学会自然言語処理研究会(NL-163-1), pp.1-9.

[2] Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web" In Proceedings of the 14th international World Wide Web conference (WWW-2005), May 10-14, 2005.

[3] Nitin Jindal and Bing Liu. "Mining Comprative Sentences and Relations." In Proceedings of 21st National Conference on Artificial Intellgience (AAAI-2006), July 16-20, 2006.

[4] L.Page,S.Brin,R.Motwani,T.Winograd. "The PageRank Citation Ranking: Bringing Order to the Web".

[5] T. Fuchi, and S. Takagi, "Japanese Morphological Analyzer using Word Co-occurence-JTAG", COLING-ACL, pp.409-413, 1998.

倉島 健　Takeshi KURASHIMA
NTT サイバーソリューション研究所 所属．2006 年京都大学大学院情報学研究科博士前期課程修了．同年，日本電信電話株式会社入社．Web マイニングの研究開発に従事．日本データベース学会会員．

別所 克人　Katsuji BESSHO
NTT サイバーソリューション研究所 所属．1994 年大阪大学大学院理学研究科数学専攻博士前期課程修了．同年，日本電信電話株式会社入社．自然言語処理の研究，ポータルサービスシステム の研究開発に従事．情報処理学会，電子情報通信学会，言語処理学会会員．

戸田 浩之　Hiroyuki TODA
NTT サイバーソリューション研究所 所属．1999 年 名古屋大学大学院 工学研究科 材料プロセス工学専攻 博士課程前期課程修了．同年日本電信電話株式会社に入社．以来，情報検索，情報抽出，Web マイニングの研究開発に従事．筑波大学大学院システム情報工学研究科博士課程に在学中．情報処理学会，日本データベース学会，ACM SIGIR，各会員．

内山 俊郎　Toshio UCHIYAMA
NTT サイバーソリューション研究所 主任研究員．1989 年東京工業大学大学院修士課程了．同年，株式会社ＮＴＴデータ入社．画像・Web データマイニング，分光色再現の研究に従事．平 18 より NTT サイバーソリューション研究所所属．

片岡 良治　Ryoji KATAOKA
NTT サイバーソリューション研究所 主幹研究員．1987 年千葉大学大学院電子工学専攻修士課程修了．同年，日本電信電話株式会社入社．トランザクションの並行処理制御方式の研究，マルチメディア情報システムの研究，ポータルサービスシステムの研究開発に従事．情報処理学会会員．

奥 雅博　Masahiro OKU
NTT サイバーソリューション研究所 主幹研究員．1984 年大阪府立大学大学院工学研究科博士前期課程修了．同年，日本電信電話公社(現 NTT)入社．機械翻訳，日本文推敲支援技術等の自然言語処理，検索をはじめとするポータルサービスの研究開発に従事．博士(工学)．情報処理学会，電子情報通信学会，言語処理学会会員．