D-005

テキスト印象マイニングに基づく質問応答システムの提案
A Web Question Answering System Based on Impression Mining from Text

熊本 忠彦 †　　**田中 克己** †,††
Tadahiko Kumamoto　　Katsumi Tanaka

1. まえがき

我々は，テキストから書き手の印象を抽出する手法を用いて，「○○ってどうなの？」型の質問（意見要求型質問）に対する回答（人々の意見）を Web 上で収集し，その要約や傾向，偏りを視覚的に提示する質問応答システムを開発している．本稿では，その進捗報告として，「怒ってばかりの母親ってどうなの？」といった意見要求型の質問に対し，関連する Web ページを集め，それぞれの印象を 2 つの印象平面（「期待⇔驚き×受容⇔嫌悪」,「喜び⇔悲しみ×恐れ⇔怒り」）上にプロットするシステムを提案する．

2. 関連研究

大量の文書データから目的のデータを効率的に探し出すための手段として，数多くの質問応答システムが提案されている [1, 2]．しかしながら，従来の質問応答システムの対話スタイルは，正解の存在を前提としているため，正解が存在しない意見要求型の質問には対応できない．意見要求型の質問に対しては，どのような意見が存在するのか，そのバリエーションを示したり，意見全体の傾向や要約を示したりすることが重要と言え，新たな対話スタイルの導入が必要とされる．

特定の製品やサービスに関する意見を Web 上で得ることを目的として，意見マイニングに関する研究 [3, 4, 5] が行われている．しかしながら，必要とされる知識（辞書）は人手による作成であり，対象とするドメイン毎に必要なため，質問の対象となる事象・事物に制限の無い場合には利用できない．一方，より汎用的な利用を前提とする研究として，評判分析に関する研究[6] がある．しかしながら，その目的は，入力となる文書データ（例えば書籍や映画のレビュー）の極性（肯定/否定，好評/不評）を判定するところまでであり，意見のバリエーションを示すことはできない．

3. テキストからの印象マイニング

テキストの印象は，4 つの印象尺度「期待⇔驚き」,「受容⇔嫌悪」,「喜び⇔悲しみ」,「恐れ⇔怒り」に対する評価値（0〜1 の実数値）を要素とする 4 次元のベクトル（印象ベクトルと呼ぶ）として記述され，テキスト中の名詞（形式名詞，副詞的名詞，数詞を除く），形容詞，動詞，未定義語の印象尺度値と重みを印象辞書から取得し，計算式に当てはめることによって求められる．

印象辞書は，文献 [7] の手法を用いて，日経新聞全文記事データベース [8]（1990 年版〜2001 年版，200 万強の記事）から自動構築された．文献 [7] では，印象尺度を構成する印象語は 1 語に限られていたが，これを複数語に拡張し，ある単語 w が 2 つの印象語群のどちらとより共起しやすいかを定式化した．この共起のしやすさを印象の強さあるいは程度と捉え，印象尺度左側の印象語群と共起しやすい場合は，その印象尺度値として 1 に近い値をとり，右側の印象語群と共起しやすい場合は，0 に近い値をとるように設計された．表 1 に今回採用された印象尺度と各印象尺度を構成する印象語を示し，表 2 に印象辞書の一部を示す．表中，各見出し語に対し，上段が印象尺度値を表し，下段が重みを表す．

表 1: 印象尺度を構成する印象語の種類

印象尺度	印象語群
期待⇔驚き	期待（する），予期（する），予想（する），期する ⇔ 驚き，驚く，びっくり（する），驚愕（する），感嘆（する），仰天（する）
受容⇔嫌悪	承知（する），了解（する），了承（する），受け入れ（る） ⇔ 嫌悪（する），嫌う，嫌いだ，嫌だ，毛嫌い（する），忌避（する）
喜び⇔悲しみ	喜び，喜ぶ，うれしい，嬉しい，楽しい，楽しむ，楽しみだ，祝福（する） ⇔ 悲しい，悲しむ，悲しみ，哀しい，哀しみ，悲哀
恐れ⇔怒り	恐れ（る），怖がる，怖い，危惧（する），怯える，恐怖（する） ⇔ 怒り，怒る，憤り，憤る，激怒（する），怒らせる，立腹（する）

表 2: 印象辞書のエントリー例

見出し語	期待⇔驚き	受容⇔嫌悪	喜び⇔悲しみ	恐れ⇔怒り
怒る	0.107	0.170	0.274	0.021
	1.304	1.179	1.300	1.622
育児	0.285	0.336	0.604	0.404
	1.346	1.199	1.273	1.105
におい	0.133	0.098	0.485	0.469
	1.304	1.205	1.309	1.113
きつい	0.397	0.190	0.575	0.422
	1.489	1.221	1.270	1.159

入力テキスト $TEXT$ は，日本語形態素解析システム Juman[9] によって単語列に分解され，名詞（形式名詞，副詞的名詞，数詞を除く），形容詞，動詞，未定義語（カタカナ，アルファベット）が抽出される．そして，単語 w の印象尺度 $e_1 \Leftrightarrow e_2$ における値 $S_{e_1 \Leftrightarrow e_2}(w)$ と重み $M_{e_1 \Leftrightarrow e_2}(w)$ が印象辞書から求められ，以下の式を用いて $TEXT$ の印象尺度値 $O_{e_1 \Leftrightarrow e_2}(TEXT)$ が算出される．

† 独立行政法人情報通信研究機構 けいはんな情報通信融合研究センター メディアインタラクショングループ
†† 京都大学 大学院 情報学研究科

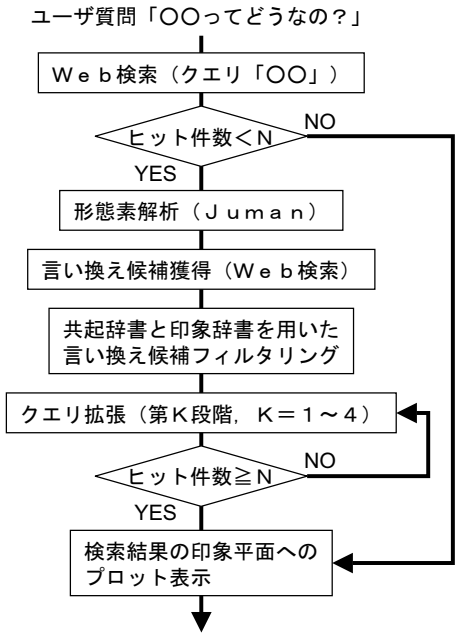

図 1: 意見要求型質問に対する回答手順

表 3: ユーザ質問として用いた例文（○○の部分を抜粋）

例文 1	怒ってばかりの母親
例文 2	育児に参加しない父親
例文 3	イラクへの自衛隊派遣
例文 4	ゴールデンウィークに海外旅行をする
例文 5	においのきつい整髪料
例文 6	幼児にお年玉をあげる
例文 7	性格の明るい幽霊
例文 8	海外旅行のために学校を休ませる
例文 9	間が悪い人
例文 10	できちゃった結婚

$$O = \sum^{TEXT} S \times |2S-1| \times M \Big/ \sum^{TEXT} |2S-1| \times M$$

この処理をすべての印象尺度に対して行い，印象ベクトル ($O_{\text{期待⇔驚き}}, O_{\text{受容⇔嫌悪}}, O_{\text{喜び⇔悲しみ}}, O_{\text{恐れ⇔怒り}}$) が生成される．なお，$|2S-1|$ 項は，傾斜配分であり，印象尺度と関係のない一般的な単語（印象尺度値は 0.5 に近い値をとる）が O 式の平均操作に及ぼす影響を抑制するために導入された．

4. テキスト印象に基づく質問応答

図 1 に意見要求型質問「○○ってどうなの？」に対する回答手順を示す．以下，表 3 に示した例文を用いて，提案システムがどのように動作するか，その特徴を示す．

[Web 検索ステージ]

提案システムは，ユーザの質問文から○○の部分を抽出し，その文字列をクエリとして Google[10] 上で Web 検索する．例えば，例文 1「怒ってばかりの母親ってどうなの？」からは，文字列「怒ってばかりの母親」を抽出し，ダブルクォーテーションで囲んで Web 検索のためのクエリとする．この初期クエリに対するヒット件数が N より小さければ，言い換え処理に基づくクエリ拡張を行い，そうでなければ，最終ステージである検索結果の印象平面へのプロット表示に進む．なお，N は，最低照会ページ数を意味し，ユーザが設定できる．

[言い換え候補獲得ステージ]

日本語形態素解析システム Juman を用いて，文字列であるクエリを単語の列に分解する*．次に，クエリ中の普通名詞，サ変名詞，形容詞，動詞，カタカナを言い換えの対象語とし，初期クエリから各対象語を取り除いた文字列を言い換え候補獲得のためのクエリとして，Web 検索を行う．例えば，初期クエリ「”怒ってばかりの母親”」に対しては,「怒る（動詞）」と「母親（普通名詞）」が対象語となり,「”ばかりの母親”」,「”怒ってばかりの”」という 2 つのクエリが生成され，Web 上で検索される．その結果，対象語のあった場所に来る単語列がそれぞれの対象語に対する言い換えの候補語として扱われる．なお，候補語の数が多いと，クエリ拡張処理によって生成されるクエリの数が増大し，処理時間も長くなるので，本稿では，対象語 1 語に対する候補語の数を 10 個以下に制限した．

[言い換え候補フィルタリングステージ]

言い換えの妥当性を共起辞書と印象辞書を用いて判定する．

共起辞書も，印象辞書と同様，日経新聞全文記事データベース（1990～2001 年版）を解析することにより構築される．まず，記事中の普通名詞，サ変名詞，カタカナを見出し語とし，各見出し語の直前／直後にある普通名詞，サ変名詞，カタカナを前接関係／後接関係として抽出し，直前／直後にある動詞，形容詞を述語関係として抽出する．また，記事中の動詞，形容詞に対しては，それぞれの直前／直後にある普通名詞，サ変名詞，カタカナを前接関係として抽出する．以上のようにして抽出された共起関係（前接関係，後接関係，述語関係）を見出し語毎にまとめ，出現頻度とともに共起辞書に登録する．

以上の方法で構築された共起辞書を用いて，言い換えの妥当性を判定する．すなわち，対象語と候補語の前接関係，後接関係，述語関係をベクトル†で記述し，以下の式を用いてコサイン類似度を求め，0.13 以上のときは，可，そうでないときは，不可と判定した．

$$\text{類似度} = \frac{\text{対象語と候補語の共通項の積の和}}{\text{対象語の大きさと候補語の大きさの積}}$$

ここで，表 3 に示した 10 個の例文に対するフィルタリングの結果の一部を表 4 に示す．

可と判定された候補語は，印象辞書を用いて，更にその妥当性を判定される．すなわち，対象語と候補語から生成される印象ベクトル間のユークリッド距離を求め，0.36 以下のときは，可，そうでないときは，不可と判定した．表 5 にフィルタリングの結果の一部を示す．

*実際には，単に分解するだけでなく，いくつかの変形操作を行う．例えば,「削除する」は「削除（サ変名詞）」と「する（動詞）」の 2 語からなるが,「削除する（動詞）」に変形し，1 語として扱う．同様に,「楽しくない（形容詞）」や「削除しない（動詞）」も 1 語として扱う．なお，この変形操作は，印象辞書構築時や印象ベクトル生成時にも行われている．

†各要素の値は，対応する共起関係の出現頻度を表す．

表 4: 共起辞書を用いた言い換え候補フィルタリングの例

（a）言い換え可（**対象語**／候補語（類似度）・・・）

母親／お母さん（0.59）・男（0.42）・自分（0.38）・私（0.33）・上司（0.24）・母（0.24）・ママ（0.21）・人生（0.19），**父親**／父（0.48）・お父さん（0.39）・男性（0.34）・夫（0.33）・子供（0.33）・男（0.31）・子育て（0.24）・状態（0.21）・パパ（0.17），**怒る**／叱る（0.19）・頷く（0.18）・働く（0.18），**育児**／子育て（0.16），**におい**／匂い（0.82）・香り（0.68）・ニオイ（0.51）・臭い（0.21）・香料（0.15）

（b）言い換え不可（**対象語**／候補語（類似度）・・・）

母親／満点（0.12）・ストレス（0.09），**父親**／具体（0.06）・里（0.06），**怒る**／言い争う（0.08）・欲張る（0.05）・出産（0.03）・喧嘩（0.03）・涙（0.02）・子育て（0.02）・発育（0.01）・比較（0.00），**育児**／家庭（0.06）・会話（0.05）・活動（0.05）・積極（0.04）・地域（0.04）・行事（0.03）・教育（0.03）・学校（0.02），**におい**／油（0.03）

表 5: 印象辞書を用いた言い換え候補フィルタリングの例

（a）言い換え可（**対象語**／候補語（距離）・・・）

母親／男（0.01）・人生（0.02）・母（0.08）・お母さん（0.08）・私（0.12）・上司（0.16）・自分（0.16）・ママ（0.18），**父親**／父（0.04）・男（0.06）・お父さん（0.11）・夫（0.15）・男性（0.19）・子供（0.21）・パパ（0.25），**怒る**／叱る（0.33），**育児**／子育て（0.08），**におい**／臭い（0.19）・香り（0.20）・匂い（0.30）

（b）言い換え不可（**対象語**／候補語（距離）・・・）

父親／子育て（0.36）・状態（0.56），**怒る**／働く（0.51）・頷く（0.72），**におい**／ニオイ（0.68）・香料（0.77）

なお，候補語が複数の単語からなる場合は，各単語と対象語との類似度／距離を算出し，その最大値／最小値を対象語と候補語の類似度／距離とした．

[クエリ拡張ステージ]

クエリ拡張処理は，**第 1 段階**（初期クエリ中の 0 個以上の対象語を候補語と言い換え，新たなクエリとする），**第 2 段階**（各クエリの連体修飾節を切り離す），**第 3 段階**（各クエリを文節に分解する），**第 4 段階**（各クエリから接続助詞を取り除く）の 4 段階であり，ヒット件数が N 以上となった時点で打ち切られ，最終ステージへと進む．

[印象平面へのプロット表示ステージ]

最終的にヒットした Web ページのうち，上位 L 件の印象尺度値が印象平面上にプロットされる．但し，L は，最大表示ページ数を意味し，ユーザが設定できる．また，各 Web ページの印象尺度値は，その Web ページが検索されたときに得られる情報（タイトルと 3 行サマリ）から求められる．

図 2 および図 3 に例文 1「怒ってばかりの母親ってどうなの？」に対するプロット表示（$L = 1000$）を示す．

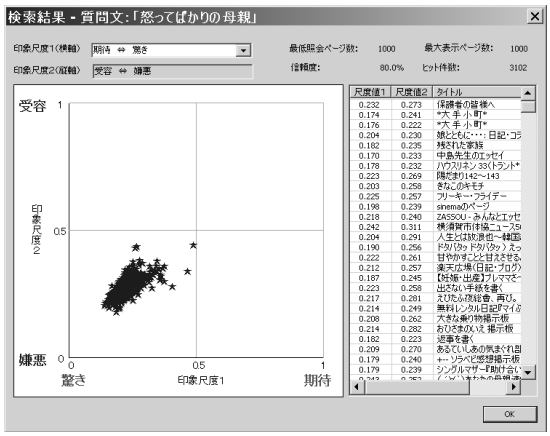

図 2: 印象平面「期待⇔驚き×受容⇔嫌悪」

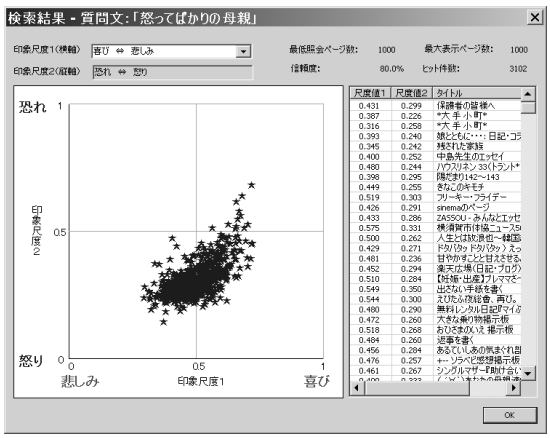

図 3: 印象平面「喜び⇔悲しみ×恐れ⇔怒り」

画面右側には各プロットの座標値とタイトルが表示されており，各行をクリックするか，各プロットをクリックすると，当該 Web ページが開くようになっている．

5. 考察

提案システムの性能を主要部分ごとに評価し，その有効性について考察する．

[言い換え候補獲得・フィルタリング]

クエリ拡張が行われた例文は 7 文あり，全部で 96 語の候補語が得られた．この 96 語に対し，言い換えの妥当性を「常に可（◎），大体の場合において可（○），例文の意味において可（△），不可（×）」の 4 段階で評価したところ，表 6 のような結果が得られた．共起関係に基づくフィルタリングにおいて不可と判定された言い換えは，55 語（全候補語の 57.3%）あったが，◎もしくは○と評価されたものはなく，△と評価されたものも，「怒る／言い争う」,「育児／教育」,「怒る／喧嘩」の 3 語だけと好成績であった．また，可と判定された言い換え（41 語）のうち，×と評価されたものは 15 語あったが，うち 9 語は印象に基づくフィルタリングにおいて不可と判定されており，その有効性を示している．

[クエリ拡張・検索結果の印象平面へのプロット表示]

表 6: 言い換え候補フィルタリング
（a）共起辞書を用いたフィルタリング

	◎	○	△	×	合計
可	11	4	11	15	41
不可	0	0	3	52	55

（b）印象辞書を用いたフィルタリング

	◎	○	△	×	合計
可	10	4	9	6	29
不可	1	0	2	9	12

表 7: クエリ拡張に伴う検索精度および意見含有率の変化

	表示ページ数	検索精度	意見含有率
第 1 段階	7 (2)	100% (5)	60.0% (3)
第 2 段階	47 (12)	77.1% (27)	92.6% (25)
第 3 段階	50 (11)	53.8% (21)	95.2% (20)

クエリ拡張（第 1 段階から第 3 段階まで）に伴う検索精度（適合率）ならびに意見含有率（意見が含まれているか否か）の変化を，例文 5「においのきつい整髪料ってどうなの？」を用いて調べた‡．結果を表 7 に示す．

クエリ拡張に伴い，検索条件が緩和されるので，検索精度が下がるのは当然と言えるが，第 2 段階まで実施しても，75%を超える高い検索精度が得られている．逆の見方をすると，25%近く誤った検索結果が得られたわけだが，その原因を調べてみると,「においのきつい」という語句が「整髪料」とは関係のない文脈（例えば「においのきつい食材を使わない」）で使われているためであった．第 3 段階の処理（文節への分解）により，検索精度が 50%にまで低下する原因も同様であった．したがって，文字列を分解して検索する際には，キーワード間の近接性を保ったまま検索する，あるいは検索結果を近接性基準でフィルタリングするといった対処が必要と考えられる．一方，意見含有率に対しては比較的よい成績が得られている.「○○ってどうなの？」という質問設定が意見収集において効果的に機能しているものと考えられるが,「においのきつい整髪料」というテーマ自体が意見を伴いやすい性質を持っている可能性もある．他の様々な質問文に対して調査を進める必要がある．

図 2，図 3 に示したように,「怒ってばかりの母親ってどうなの？」というクエリに対し，もっともらしい表示がなされている．しかしながら，現在のシステムは，検索結果画面に表示されるタイトルと 3 行サマリから印象を算出しているため，Web ページに含まれる意見部分の印象を正確に反映しているとは言えない．今回収集された Web ページの多くは，複数の文書を含むブログや掲示板などであるため，Web ページ本文中の意見部分を同定し，その印象を抽出することが必須と言える．今後の課題としたい．

‡パラメータを最低照会ページ数 $N = 50$，最大表示ページ数 $L = 50$ に設定し，各段階においてヒットした Web ページを調べた．このとき，整髪料のにおいに関して何らかの言及がある場合を「正解」，ない場合を「誤り」とし，また書き手の意見が明示されている場合を「意見を含む」とした.「表示ページ数」項の括弧内の数字は，すでに削除されていて／書き換えられていて，アクセスできなかった Web ページの数を示す．なお，第 3 段階における表示ページ数は L の制限によるもので，実際には 8,503 件ヒットしている．

6. まとめ

本稿では,「○○ってどうなの？」形式の質問文（意見要求型）に対する回答（人々の意見）を，Web 上で収集し，その印象を印象平面上にプロット表示するシステムを提案した．本システムは，一定数の意見を確保するため，言い換え処理に基づくクエリ拡張を行い，言い換えの妥当性を共起辞書と印象辞書を用いて判定する点に特徴がある．Web ページの印象は，処理時間の関係もあり，Web ページそのものではなく，検索結果として得られるタイトルと 3 行サマリから抽出されたが，実際に収集された Web ページの多くは，複数の文書を含むブログや掲示板などであったため，意見部分とは関係のないタイトルがついた Web ページも多かった．

今後は,（1）Web ページから意見部分を同定するための手法の開発,（2）意見全体の要約や傾向を概観するためのビューワーの開発,（3）印象マイニング手法の高精度化,（4）個人差を考慮した印象マイニングを行うためのユーザモデルの導入，等に取り組みたい．

参考文献

[1] 村田真樹，内山将夫，井佐原均，類似度に基づく推論を用いた質問応答システム，情処学自然言語処理研報，Vol.2000，No.011，pp.181-188 (2000)．

[2] 佐々木裕，磯崎秀樹，平博順，平尾努，賀沢秀人，鈴木潤，国領弘治，前田英作，SAIQA：大量文書に基づく質問応答システム，情処学自然言語処理研報，Vol.2001，No.086，pp.77-82 (2001)．

[3] 立石健二，石黒義英，福島俊一，インターネットからの評判情報検索，情処学自然言語処理研報，Vol.2001，No.069，pp.75–82 (2001)．

[4] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima, Mining Product Reputations on the Web, In Proceedings of the 8th ACM SIGKDD Conference, pp.341–349, Edmonton, Alberta, Canada (2002).

[5] 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，情処学自然言語処理研報，Vol.2004，No.093，pp.1-8 (2004)．

[6] Peter D. Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, Proc. of the Conference on Association for Computational Linguistics, Philadelphia, USA (2002).

[7] 熊本忠彦，田中克己，Web ニュース記事を対象とする喜怒哀楽抽出システム，インタラクション 2005，Vol.2005，No.4(A-103)，pp.25–26 (2005)．

[8] 日経全文記事データベース DVD-ROM 版，1990–1995，1996–2000，2001 年版，日本経済新聞社．

[9] 黒橋禎夫，河原大輔，日本語形態素解析システム JUMAN version 4.0 (2003)．

[10] Google, http://www.google.co.jp/