問合せの主観性にロバストな商品検索システム

杉木 健二† 松原 茂樹‡

†名古屋大学大学院情報科学研究科 ‡名古屋大学情報連携基盤センター

1 はじめに

近年，インターネット利用者の増加に伴い，楽天やAmazon，yahoo!など，EC(e-commerce) サイトが急速に増加している．これらのサイトを効率的に利用するために商品の検索が不可欠である．従来の商品検索システムでは，商品カテゴリによる検索，商品名，型番などのキーワードによる検索，性能や価格帯など，データ項目の選択による検索などの機能を備えている．

しかし，消費者のニーズは多様であり，従来の検索システムによってこのような多様性に対応することは容易ではない．特に，消費者の主観的な要求に対応した検索を実現することは困難である．

本稿では，自然言語入力に基づく商品検索システムを提案する．本研究では，問合せの主観性に対応するために商品のレビューを用いる．本研究では，商品レビューやサービスに対する意見，感想などを，**意見テキスト**と呼ぶ．

提案するシステムでは，意見テキストから**意見情報**（意見対象，項目，値）を抽出し，検索クエリから**要求情報**（要求対象，項目，値）を取り出す．それらの一致率に基づいて，要求情報と意見情報を対応付ける．

意見情報を高い精度で抽出するために，従来研究では半自動的に辞書を構築する手法が提案されている [1, 2, 3]．しかし，辞書を構築するコストは高く，ドメインごとに作成する必要があるといった問題がある．本システムでは，大規模かつ広範囲のドメインでの処理を想定しているため，辞書を構築せずに意見情報を抽出する．

2 意見情報

本研究では意見情報を，「製品やサービスに対する評価や意見，事実，性質を表している情報」と定義する．また，消費者が欲しい商品とは，「色は赤で，サイズは軽量で，デザインはシンプルで，…」といったように，商品のある特徴 (項目) とその値（もしくは制約）の組で示されることが多い．そこで，本稿では，項目と値の組に，意見対象 (商品名) を加えた 3 項組を，意見情報 (意見対象，項目，値) と定義する．

MP3 プレイヤー（プレイヤー A）の意見テキスト例を図 1 に示す．図 1 のテキストから，

- （プレイヤー A，デザイン，汚れが目立つ）
- （プレイヤー A，デザイン，シンプル）
- （プレイヤー A，デザイン，良い）
- （プレイヤー A，充電，早い）

> 鏡面デザインは汚れが目立つので好きではありませんが，シンプルで良いと思います．バッテリは，まだ長時間使用していないので何ともいえませんが，充電が早いのは助かります．操作性ですが，慣れれば問題ありませんが，曲のサーチの仕方はイマイチで，今何を操作してどう探しているのかがわかりにくいインターフェイスになっているので，改善の余地はあるでしょう．

図 1: MP3 プレイヤーの意見テキスト例

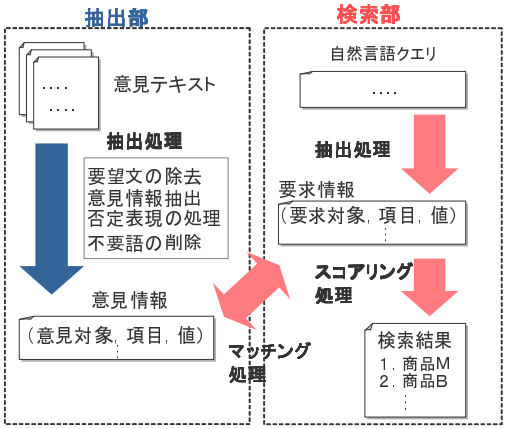

図 2: 本システムの構成

- （プレイヤー A，操作性，問題ない）

といった意見情報を抽出できる．これらの意見情報から，「デザインがシンプルで，充電が早い MP3 プレイヤー」といった問合せに該当することが分かる．

3 商品検索システム

システムの概要を図 2 に示す．本システムは，抽出部と検索部からなる．

抽出部では，意見テキストから意見情報を抽出する．意見テキストを係り受け解析した結果に変換ルールを適用することにより，意見情報を抽出する．

検索部では，ユーザが入力した問い合わせに合致した商品を出力する．問い合わせとして自然言語による記述を許容する．問い合わせから生成した要求情報を意見テキストから抽出した意見情報と対応付ける．

4 意見情報の抽出

4.1 意見情報への変換ルール

意見情報として，ある意見対象に関する項目と値の組を意見テキストから抽出する．項目と値の関係の多く

は，意見テキスト中では，主語と述語の関係，被修飾・修飾の関係として出現する．

- 主語・述語の関係
  「料理がおいしい」⇒（Aホテル，料理，おいしい）
- 修飾・被修飾関係
  「きれいな部屋」⇒（Bホテル，部屋，きれい）

上述の関係について予備調査を行ったところ，以下のような係り受けパターン例[1]が多く確認された．

(1) 部屋が $_X$ →きれい $_Y$

(2) きれいな $_Y$ →部屋です $_X$

(3) 部屋の $_{X_1}$ →ベッドが $_{X_2}$ →快適です $_Y$

(4) 部屋が $_X$ →きれいで $_{Y_1}$ → 快適でした $_{Y_2}$

上記のパターン例が A ホテルに対する意見テキストに出現した場合，以下のような意見情報を抽出すればよい．

(1) （A ホテル，部屋 $_X$，きれい $_Y$）

(2) （A ホテル，部屋 $_X$，きれい $_Y$）

(3) （A ホテル，部屋 $_{X_1}$-ベッド $_{X_2}$，快適 $_Y$）

(4) （A ホテル，部屋 $_X$，きれい $_{Y_1}$），
（A ホテル，部屋 $_X$，快適 $_{Y_2}$）

パターン (3) では，X_1 と X_2 を連結して項目とし，また，(4) では，Y_1 と Y_2 のそれぞれについて，意見情報を作成する．以上の観察から，本研究では，以下の変換ルールを使用する．ここで O は意見対象を示す．

(1) $X \rightarrow Y \Rightarrow (O,X,Y)$

(2) $Y \rightarrow X \Rightarrow (O,X,Y)$

(3) $X_1 \rightarrow X_2 \rightarrow Y \Rightarrow (O,X_1-X_2,Y)$

(4) $X \rightarrow Y_1 \rightarrow Y_2 \Rightarrow (O,X,Y_1),\ (O,X,Y_2)$

ルール通用の汎用性を高めるために，さらに，(3),(4) を組み合わせたルール：

(5) $X_1 \rightarrow X_2 \rightarrow Y_1 \rightarrow Y_2 \Rightarrow (O,X_1-X_2,Y_1),(O,X_1-X_2,Y_2)$
例：浴室の $_{X_1}$ →浴槽が $_{X_2}$ →広くて $_{Y_1}$ → くつろげました $_{Y_2}$ ⇒ (A ホテル，浴室-浴槽，広い)，　(A ホテル，浴室-浴槽，くつろげる)

また，項目がなく値のみ存在する場合に対応するためのルール：

(6) $Y \Rightarrow (O,\ -\ ,\ Y)$
例：親切でよかったです．また来たいと思います．⇒ (A ホテル，- ，親切)，(A ホテル，- ，よい)
（ハイフン (-) は，要素が存在しないことをを示す）

を設けた．

4.2 意見情報抽出の手順

4.2.1 前処理

意見テキストを文単位に分割し，各文を係り受け解析することにより抽出の前処理とする．ただし，要望文については以下のように除去する．

[1] X,Y は，それぞれ項目，値を含む文節を表し，矢印「→」はそれぞれの係り受け関係を表す．

要望文とは，「〜して欲しい」「〜と望ましい」「〜ば嬉しい」などの期待や願望，依頼などの要望表現を含む文である．要望文であるかどうかの判定は，要望表現が含まれているかにより行った．「仮定形」や「命令形」の形態素が含まれる文，もしくは，「のに」，「たら」，「たい」などの形態素が含まれる文を要望文とした．

4.2.2 抽出処理

抽出処理では，上述した次の 6 つの変換ルールを適用して，意見テキストから意見情報を抽出する．意見テキストにこれらのルールを適用するための制約として，以下の制約を与えた．

ルール (1) X：「名詞+は/が/も」，Y：動詞，形容詞，名詞 (サ変名詞＋する)

ルール (2) X：「名詞+は/が/も/を/に/だ/です」，Y：形容詞（接続助詞を含まない）

ルール (3) X_1：「名詞＋の (助詞)」，X_2：「名詞+は/が/も」，Y：動詞，形容詞，名詞 (サ変名詞＋する)

ルール (4) X：「名詞+は/が/も」，Y_1，Y_2：動詞，形容詞，名詞 (サ変名詞＋する)

ルール (5) X_1：「名詞＋の (助詞)」，X_2：「名詞+は/が/も」，Y_1，Y_2：動詞，形容詞，名詞 (サ変名詞＋する)

ルール (6) Y：動詞，形容詞，名詞 (サ変名詞＋する)

(1),(3),(4),(5) のルールは項目と値が主語・述語関係となる場合，(2) のルールは項目と値が被修飾・修飾関係となる場合，(6) のルールは項目が存在しなかった場合に適用する．まず，意見テキスト中の主語・述語関係を判定し，(1),(3),(4),(5) のルールの適用を試みて，これとは独立に，修飾・被修飾関係を判定し，(2) のルールを適用を試みる．ルールが適用されない値候補があれば，最後に (6) のルールの適用を試みる．

抽出処理例を図 3 に示す．意見テキストを係り受け解析し，項目候補と値候補を特定する．これらの項目候補と値候補に対して，(5),(3),(4),(1) の順で適用可能性を調べる．このテキストでは，(4),(5) のルールを適用できる．また，ルールが適用されていない値候補があるので，この値候補に対して (6) のルールを適用し，最終的に 5 つの意見情報が抽出される．

4.2.3 後処理

否定表現処理では，値に否定表現が含まれる場合には表現を統一する．この処理により，例えば「十分でない」「十分ではありません」「不十分だ」という表現を同一の値として扱うことができる．

項目として用いる形態素を名詞のみとし，値として用いる形態素を名詞，形容詞，動詞，副詞，接頭辞，接尾辞とする．接頭辞，接尾辞は，「大満足」，「１万円以下」などの表現を扱うために除去しない．また，「方」，「こと」，「もの」，「思う」，「考える」など，重要な意味を形成しない形態素を除去する．不要語を除去することにより，複数の表現を同一の表現として検索することができる．

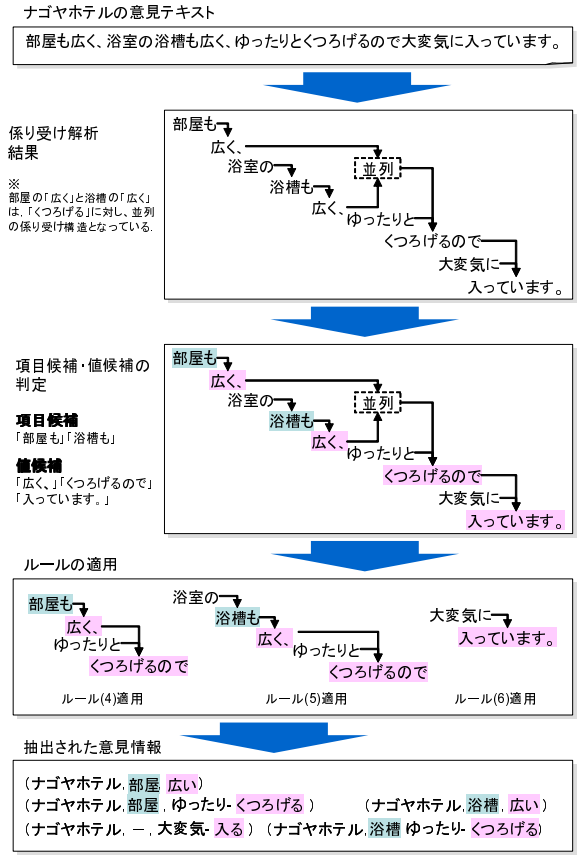

図 3: 抽出処理の例

5 商品の検索

要求情報とは（要求対象，項目，値）の 3 項組である．このうち，項目および値は，意見情報と同様である．一方，要求対象とは，商品や宿泊施設などのクラス（もしくは，商品カテゴリ）であり,「ノートパソコン」,「車」,「ホテル」などが該当する．

検索処理では，まず，検索クエリに対して，意見情報の抽出処理と同様の処理を行い，項目と値の組を抽出する．要求対象については，商品クラスに該当する名詞句を抽出する．ただし，本稿では，検索対象として宿泊施設を対象としているので，要求対象を「ホテル」に固定した．

次に，これらの要求情報と意見対象を一致率に基づいて対応付ける．一致率を用いることにより，部分一致による対応付けが可能となる．最後に，各意見対象のスコアを計算し，スコアの高い順に検索対象を表示する．

5.1 要求情報と意見情報との対応付け

本研究では,「料理はおいしくて，部屋が広くて，値段が安い，…」のように，検索クエリが項目と値の組で入力されると仮定している．したがって，抽出部と同様の処理をすることによって，クエリから要求情報を抽出できる．例えば,「料理がおいしくて，部屋がきれいなホテルを探しています.」といった検索クエリが与えられた場合，(ホテル，料理，おいしい), (ホテル，部屋，きれい）という要求情報を抽出できる．これらの要求情報に対して，合致する意見情報を検索し，これらの意見情報の意見対象を提示すればよい．

$$c\_rate(i,j) = \begin{cases} \frac{EOP_{match}}{EOP_{all}} \times \frac{EV_{match}}{EV_{all}} & (\text{項目がある}) \\ 0.1 \times \frac{EV_{match}}{EV_{all}} & (\text{項目がない}) \\ 0 & (otherwise) \end{cases}$$

EOP_{match} : 要求情報の項目中の形態素一致数
EOP_{all} : 要求情報の項目中の全形態素数
EV_{match} : 要求情報の値中の形態素一致数
EV_{all} : 要求情報の値中の全形態素数

図 4: 一致率の計算

$$Score(Q,O_i) = \sum_{j \in Q} PF_{ij} \cdot IOF_j$$

$$IOF_j = \log(\frac{1}{of_j + 1} + 1)$$

$$PF_{ij} = \sum_{k_i \in O_i} pf_{k_i} \times c\_rate(j,k_i)$$

of_j: クエリ中の要求情報 j との一致率が 0 より大きい意見情報が出現する意見対象数
pf_{k_i}: 意見情報 k が意見対象 i に出現する頻度

図 5: スコア計算式

本研究では，要求対象を「ホテル」に限定しているので，要求情報と意見情報との対応付けは，これらの情報の項目と値の組との対応付けにより行う．

本研究では，要求情報と意見情報とを対応付けるために，一致率を定義した．図 4 に一致率の計算式を示す．要求情報の項目 i と意見情報 j との一致率は，要求情報の項目中の全形態素のうち意見情報の項目との形態素一致数と，要求情報の値中の全形態素のうち意見情報の値との形態素一致数との積により計算される．一致率に基づく対応付けをすることにより，要求情報と意見情報との部分一致による対応づけができ，さらに，意見情報が要求情報と一致しているほどスコアが高くなると期待できる．例えば，要求情報 a（ホテル，料理，おいしい）と意見情報 b（Aホテル，料理，とても-おいしい）があった場合，一致率 $c\_rate(a,b) = 1 \times 1 = 1$ となる．また，要求情報 c（ホテル，料理，とても-おいしい）と意見情報 d（Bホテル，料理，おいしい) の場合，一致率は，$c\_rate(c,d) = 1 \times \frac{1}{2} = \frac{1}{2}$ となる．このとき，要求情報 c と意見情報 b の一致率は，$c\_rate(c,b) = 1 \times \frac{2}{2} = 1$ となり，要求情報中の副詞が意見情報中に含まれていれば，より一致率が高くなる．

また，項目がない要求情報が検索される場合がある．このような要求情報に対して，同様に項目がない意見情報とを対応付ける．一致率の項目の部分が計算できないので，図 4 に示すように，要求情報中の値の全形態素における形態素一致数に 0.1 を掛けることにより，一致率を計算している．0.1 と小さな値を設定するのは，項目と値が含まれる要求情報に対する影響を小さくするためである．

表 2: 実験結果

クエリ	1	2	3	4	5	6	7	8	9	10
評価平均	4.00	3.80	1.89	2.40	3.10	1.00	3.30	2.60	4.00	3.00
システム平均	5.09	7.93	0.09	0.04	1.34	5.57	43.47	1.94	9.04	0.51

表 1: 実験に用いた 10 個の検索クエリ

(1)	部屋にパソコンがあるか、自分のパソコンでインターネットが使えるホテル
(2)	ぐっすり眠れる部屋で、ベッドの寝心地が良くて、防音対策がいいホテル
(3)	少し非日常的なロマンチックなお風呂が付いているホテル
(4)	一人 5000 円以下のホテル
(5)	遅い時間にチェックインとチェックアウトができるホテル
(6)	連続で泊まると宿泊の割引があるホテル
(7)	部屋が綺麗で、タバコ臭くないホテル
(8)	無料の朝食バイキングが付いているホテル
(9)	駅に近い、もしくは、無料送迎バスがあるホテル
(10)	ホテルの近くに、遊べるところがたくさんあるホテル

5.2 スコアリング

各意見対象に対するスコアの計算方法を図 5 に示す．PF_{ij} は，上述した一致率とある意見対象における意見情報の出現頻度との積を加えることにより求まる．意見対象 O_i の完全に一致した（一致率 $= 1$）意見情報に加え，部分一致（一致率 < 0）の意見情報の出現頻度を考慮することができる．完全一致のみでスコアリングを行った結果の場合と比較して，再現率が高くなることを期待している．

計算式の $PF_{ij}, IOFj$ は，$tf \cdot idf$ 計算式のそれぞれ tf, idf に相当し，$tf \cdot idf$ に類似した計算式を用いた．PF_{ij} は各意見対象における（擬似的な）頻度を示している．また，IOF_j は，要求情報の出現対象数の逆数の対数である．これは，ある要求情報に対して，出現する意見対象が少ないほど価値がある情報であり，意見対象中の出現頻度が大きいほど，意見対象 O_i のスコアが高くなるように設定している．つまり，要求情報と合致する意見情報の出現頻度が大きいほど，意見対象が適合している可能性が高く，要求情報が合致する意見対象数が少なければ，その要求情報は貴重であり，それに合致する意見情報を含む意見対象であれば，その意見対象がクエリにより適合している可能性が高いとする．

6 検索実験

6.1 実験方法

宿泊施設の予約サイト[2] から，意見テキストを取得した．実験には，ホテル 1,670 件に対する意見テキスト 265,527 件を用いた．係り受け解析器として KNP[4] を用いた．

実験では，図 1 に示す 10 個の検索クエリを被験者が作成し，検索結果の妥当性を検証した．検索結果上位 10 件を使用し，被験者がクエリとの一致度を 4 段階で評価した．評価基準を以下に示す．

4 ：商品が検索クエリとほとんど一致

3 ：商品が検索クエリと部分的に一致，関連している

2 ：商品が検索クエリと多少関連している

1 ：商品が検索クエリとまったく関連していない

被験者は，各ホテルのクエリに関連した意見テキストを閲覧し，判定する．

6.2 実験結果

表 1 の検索クエリに対する実験結果を表 2 に示す．表 2 の各クエリの番号は，表 1 の番号と一致している．表 2 は，10 個の各クエリに対する被験者の 4 段階評価の平均と，検索スコアの平均をそれぞれ示している．

6.3 考察

表 2 の実験結果から，スコアが 5 以上（クエリ (6) 以外）であれば，被験者の評価が良いということが分かった．また，システムのスコアが 2 以下であれば，被験者の評価があまり高くないということが分った．

被験者の評価が低くなった原因として，一致率の問題が挙げられる．項目と値それぞれの形態素が部分一致すればよいということと，形態素末が一致が必須であるという問題である．前者の場合は，ノイズが多い結果となってしまい，後者の場合は，再現率が低下する原因となった

また，クエリ 6「連続で泊まると宿泊の割引があるホテル」のような条件付きの検索はほとんど行えないということが分かった．

7 おわりに

本稿では，ロバストな商品検索の実現を目的とし，曖昧で主観的な自然言語入力による問合わせに対応した検索システムを提案した．問合わせの主観性に対応するために，商品について書かれた意見テキストを使用した．ホテルを対象に検索実験を行った．被験者による評価の結果，主観的で曖昧な問合せに対して，検索が可能であることを確認した．

参考文献

[1] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. テキストマイニングによる評価表現の収集. 情報処理学会研究報告, 自然言語処理研究会, 2002-NL-154, pp. 77–84, 2003.

[2] 立石健二, 福島俊一, 小林のぞみ, 高橋哲朗, 藤田篤, 乾健太郎, 松本祐治. 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. 情報処理学会研究報告, NL-163, pp. 1–8, 2004.

[3] 矢野宏美, 目良和也, 相沢輝昭. 嗜好を考慮した評判情報検索手法. 情報処理学会研究報告, NL164, pp. 165–170, 2004.

[4] 黒橋禎夫. 日本語構文解析システム knp version 2.0, 2005.

[2]楽天トラベル「お客さまの声」
http://travel.rakuten.co.jp/auto/tabimado_bbs_top.html