E-056

クエリの主観性に頑健な商品検索システム
A Product Retrieval System for Subjective Queries

杉木 健二 †
Kenji Sugiki

松原 茂樹 ‡
Shigeki Matsubara

1. まえがき

近年，インターネット利用者の増加に伴い，楽天やAmazonなど，ECサイトが急速に増加している．これらのサイトでは膨大な数の商品を扱っており，優れた検索環境を提供することが不可欠である．しかし，ECサイトの多くは，商品名や商品カテゴリなど，あらかじめ定められたデータ項目に対する検索機能を備えているに過ぎない．利用者の要求は多様であり，また，主観的であることも少なくなく，そのような要求に適応することは容易ではない．

これに対して，利用者が自然言語で要求を記述し，検索システムへの入力とすることが考えられる．これまで，自然言語インタフェースを備えた商品検索システムが提案されている (例えば，宿泊施設を対象にしたシステム [1])．しかし，自然言語クエリをデータベース言語に変換する方式では，問合せの主観性，多様性に対応することは困難である．

本論文では，自然言語で表現された主観的な要求に対して頑健に対応可能な商品検索システムを提案する．本研究では，入力された主観的な検索クエリに合致した商品を提示するために，消費者による意見を用いる．最近では，ECサイトの商品にそのような口コミ情報が提供されることは一般的であり，また，@ cosme, 価格 com などの口コミサイトや，商品のレビューが掲載されたblogなども多く存在し，消費者の商品購入時の判断材料として利用されている．本システムでは，これらの商品に関する情報を用いて，検索クエリがある商品に関する意見（以下，**意見テキスト**）と一致すれば，その商品は利用者の要求に合致しているとして検索を実行する．

本研究では，宿泊施設を対象に検索システムを作成した．宿泊者が記した情報を利用することにより,「ベッドがやわらかく，寝心地がよい宿」といった，従来のECサイトでは受け付けることができなかった主観的な検索クエリにも対応できる．約22万件の意見テキストを使用して検索実験を実施した結果，本システムの有用性を確認した．

2. 意見テキストを用いた商品検索システム

提案するシステムでは，自然言語で表現された検索クエリに対して，該当する内容が記載された意見テキストの商品を検索結果として提示する．例えば，図1左の検索クエリの入力に対して，プラズマテレビ「TH-42PX500」に図1右の意見テキストが存在すれば，クエリに合致した商品であるとみなす．

検索クエリと意見テキストとの内容の一致を判定するために，双方を意味表現に変換し，意味表現間の一致率を計算する．本研究では，検索クエリ及び意見テキスト

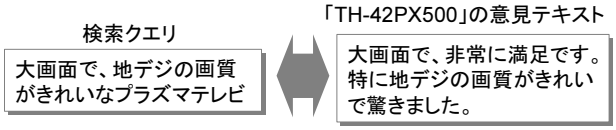

図 1: 検索クエリと対応する意見テキストの例

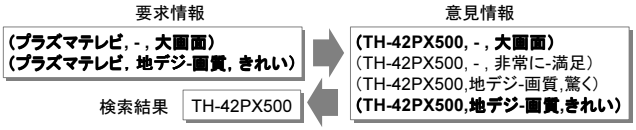

図 2: 要求情報と意見情報との対応付けの例

の情報を，“(**対象, 項目, 値**)”の3項組で表現する．ここで,「対象」は商品名や商品カテゴリなどを表し,「項目」と「値」はそれぞれおその商品の特性とその価を示す．これは，評判情報抽出に関する従来の研究 [2, 3, 4, 5] で使用された表現形式と同一であり，商品の特徴を表現するのに適している．ただし，従来研究は，評判情報の抽出が目的であり，評判に関わる形容詞をキーにした抽出手法，ならびに，評判表現辞書の構築が検討されてきた．一方，本研究では，評判などの主観的な情報だけでなく，事実などの客観的な情報も同様に抽出するため，評判表現に特化することなく，副詞，動詞，名詞等の自立語も同様に抽出の対象とする．

本研究では，意見テキストから “(**意見対象, 項目, 値**)”の3項組を抽出し，これを**意見情報**と呼ぶ．例えば,「TH-42PX500」について「画質がきれい」という意見があれば，意見情報 (TH-42PX500, 画質, きれい) を抽出する．一方，検索クエリを “(**要求対象, 項目, 値**)”の3項組に変換し，これを**要求情報**と呼ぶ．例えば,「画質がきれいなプラズマテレビ」というクエリを，要求情報 (プラズマテレビ, 画質, きれい) に変換する．この場合，この要求情報と上記の意見情報とを比較すると，プラズマテレビ「TH-42PX500」は，検索クエリに合致する商品であると判断できる．図1の検索クエリ及び意見テキストから抽出される要求情報及び意見情報，ならびにその対応付けの例を図2に示す．

システムの構成を図3に示す．システムは意見テキストから意見情報を抽出する抽出部，ならびに，検索クエリに合致する商品を検索する検索部から構成される．抽出部では，意見テキストの各文に，係り受けパターンに基づく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，クエリから要求情報を生成し，意見情報との一致度を計算し，スコア順に商品を提示する．

3. 意見情報の抽出

本節では，意見テキストから意見情報を抽出する手法について説明する．

†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

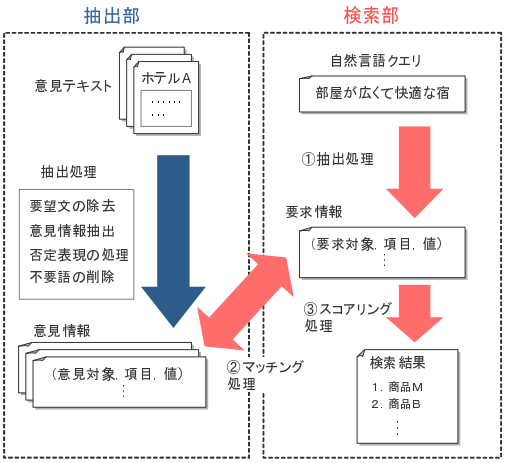

図 3: 商品検索システムの構成

3.1 意見情報への変換ルール

本研究で対象とする意見テキストは，楽天トラベルなど，特定の商品について意見しているサイトのテキストである．本研究では，意見の対象となっている商品名を取得し，項目と値の組を意見テキストの本文から抽出する．

意見テキスト中では，項目と値の組は以下の「主語と述語の関係」もしくは「被修飾・修飾の関係」として出現すると考えられる．本研究では，文節間の係り受け関係を用いることにより，これらの関係を特定する．

- 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)

変換ルールを作成するために，意見テキスト中の項目と値の出現と，係り受けパターンの出現との関係を調査した．その結果，以下の係り受けパターン例§が確認された．

(1) 部屋が $_X$ →きれい $_Y$

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

(3) きれいな $_Y$ →部屋です $_X$

これらの 3 つのパターンからそれぞれ以下のような意見情報を抽出できる．

(1) (A ホテル, 部屋 $_X$, きれい $_Y$)

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

(3) (A ホテル, 部屋 $_X$, きれい $_Y$)

以下より，次に示す変換ルールを作成できる．ここで O は意見対象を示す．

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

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

§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．

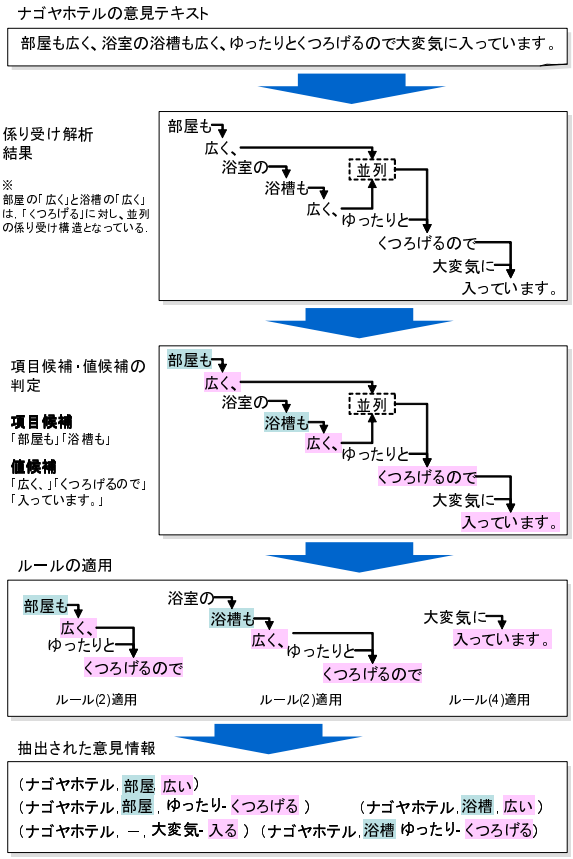

図 4: 抽出処理の例

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

また，意見テキスト中では項目部分が記述されず，値のみ出現する場合に対応するため，以下のルールを追加する．

(4) $Y \Rightarrow (O,$ - , $Y)$
例：親切でよかったです．⇒ (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）

より詳細な情報も抽出するために，これらのルールに加えて，項目の文節に係る「名詞＋の」のパターンが含まれる文節を項目に含め，値の文節に係る「副詞」や「名詞＋格助詞」のパターンなどが含まれる文節を値に含める処理をする．

3.2 意見情報抽出の手順

3.2.1 前処理

前処理では，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．さらに，「～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は，検索クエリとしてはほとんど入力されないため除去する．

3.2.2 抽出処理

抽出処理では，上述した 4 つの変換ルールを適用し，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，以下のような品詞パターンの制約を与えた．

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」，Y_1，Y_2：動詞，形容詞，名詞 (サ変名詞＋する)

ルール (3) 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 (4) Y：動詞，形容詞，名詞 (サ変名詞＋する)

これらのルールを，ルールの制約が強い順に適用する．(1),(2) のルールは項目と値が主語・述語関係となる場合に適用され，(2),(1) の順にルールを適用する．次に，(3) のルールは，項目と値が被修飾・修飾関係となる場合に適用される．最後に，値のみ存在し項目候補が存在しない場合，(4) のルールを適用する．

抽出処理例を図 4 に示す．まず，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，(1),(2),(3),(4) の順でルールを適用する．このテキストでは，(1),(2) のルールを適用できる．また，(1),(2),(3) のルールが適用されない値候補があるので，この値候補に対して (4) 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．

3.2.3 後処理

後処理として，否定表現処理と不要語処理をする．これらの処理は，検索の再現性を高める効果がある．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．

不要語処理では，特定の品詞や不要語を除去する．項目として用いる形態素を名詞のみとし，それ以外の品詞は項目から除去する．また，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．さらに,「方」,「こと」,「もの」,「思う」,「考える」など，重要な意味を形成しない語を除去する．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可能となる．

4. 商品検索

要求情報の要求対象（商品クラス，商品カテゴリ）に対応する意見情報の意見対象（商品名）をそれぞれの情報の項目と値とを対応付けることにより検索する．ここで，要求対象とは，例えば「ノートパソコン」,「車」,「ホテル」などである．ただし，本論文では単一の商品カテゴリ（宿泊施設）を対象にしており，商品がどの商品カテゴリに属するかという判定は行わない．

検索部では，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判定する．なお，本論文では，クエリの主観性に対する性能を確認するために，検索する対象は宿泊施設のみとする．すなわち，ユーザが入力するクエリを「〜宿」という形式で終了するように限定し，要求対象を「宿泊施設」に固定する．

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

EOP_{match}： 要求情報の項目中の文節一致数
EOP_{all}： 要求情報の項目中の全文節数
EV_{match}： 要求情報の値中の文節一致数
EV_{all}： 要求情報の値中の全文節数
※ただし，値は末尾の文節の一致が必須

図 5: 一致率の計算

$$Score(Q,O) = \sum_{q_j \in Q} PF_j \cdot IOF_j$$

$$PF_j = \sum_{o_i \in O} pf_i \times c\_rate(i,j)$$

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

pf_i: 意見情報 o_i が意見対象 O に出現する頻度
of_j: クエリ中の要求情報 j との一致率 > 0 である意見情報が出現する意見対象数
ON: 全意見対象数

図 6: スコア計算式

4.1 要求情報と意見情報との対応付け

抽出部と同様の抽出処理により，クエリから要求情報を抽出する．例えば,「部屋がきれいで，朝食が付いて，値段が安い宿」というクエリから，要求情報 (宿, 部屋, きれい), (宿, 朝食, 付く), (宿, 値段, 安い) が抽出される．

本研究では，要求情報と意見情報との項目と値の組を比較するために，一致率を計算する．図 5 に一致率の計算式を示す．文節単位で一致率を計算する．一致率に基づいた対応付けにより，要求情報に関連した部分一致の意見情報も対応付けることができ，さらには，意見情報が要求情報と一致しているほど，その意見情報のスコアが高くなると期待できる．

例えば，要求情報 a(宿, 対応, 満足) と意見情報 b(A ホテル，対応，いつも-満足) 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(宿, 対応, いつも-満足) と意見情報 d(B ホテル, 対応, 満足) の場合，一致率は，$c\_rate(c,d) = 1 \times \frac{1}{2} = \frac{1}{2}$ となる．

項目がない要求情報に対しては，項目がない意見情報に対応付ける．図 5 に示すように，項目に該当する部分を 0.1 と設定する．これは，再現性を持たせ，項目が含まれる要求情報に対してあまり影響を与えない値とした．

4.2 スコアリング

各商品（意見対象）のスコアを計算し，クエリにより合致した商品から順に提示する．本研究では以下のような，信頼できる，かつ，個別性の高い情報を含むような商品がより良い商品であるとする．

(1) より多くの人が同じ意見を記述していれば，その情報はより信頼できる情報である．(TF 的要素)

表 1: 実験結果

	客観的クエリ			主観的クエリ							平均
クエリ#	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	
検索結果宿数	3	9	10	10	10	10	4	10	10	10	
評価平均	4.00	2.89	3.10	3.60	4.00	3.30	4.00	2.40	1.00	4.00	3.23

客観的クエリ

クエリ#	検索クエリ
(1)	チェックアウト後も荷物を預かってもらえる宿
(2)	食事が和食と洋食で選べる宿
(3)	チェックアウト時間が遅い宿

主観的クエリ

クエリ#	検索クエリ
(4)	風呂が広くてアメニティが充実している宿
(5)	部屋の照明が明るい宿
(6)	周辺が閑静な雰囲気で，落ち着いて過ごせる宿
(7)	コンビニやレストランが近くて，食事に困らない宿
(8)	落ち着いて朝食がとれる宿
(9)	部屋のインテリアが格調高い宿
(10)	ベッドがやわらかく，寝心地がよい宿

図 7: 実験に用いた 10 個の検索クエリ

(2) ある意見が書かれた商品（意見対象）数が少なければその意見は個別性が高く貴重な情報である．(IDF 的要素)

各商品（意見対象）に対するスコアの計算方法を図 6 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 の完全に一致した意見情報に加え，関連した部分一致の意見情報も含めたスコアリングが可能である．PF_{ij} は各意見対象における頻度を表し，IOF_j は，要求情報の出現対象数の逆数を対数化したものである．PF_{ij} は，上記の (1) に該当し，IOF_j は上記の (2) に該当する．

5. 検索実験

5.1 実験方法

本システムの有効性を確認するために，宿泊施設検索システムを作成し，検索実験を行った．実験では，宿泊施設の予約サイト* から，意見テキストを取得し，1,359 件のホテルに対する意見テキスト 220,159 件を用いた．係り受け解析には KNP[6] を使用した．

評価は，図 7 に示す 10 個の検索クエリに対して，検索結果の妥当性を検証することによって実験した．検索の結果生成された上位 10 件に対して，被験者がクエリとの対応度を以下に示す 4 段階で評価した．

4 ：宿泊施設が検索クエリとほとんど適合している
3 ：宿泊施設が検索クエリと部分的に適合している
2 ：宿泊施設が検索クエリと少し適合している
1 ：宿泊施設が検索クエリと全く適合していない

検索クエリの設定及び妥当性の判定は，著者とは異なる被験者が実施した．

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

5.2 実験結果

実験結果を表 1 に示す．表 1 は，各クエリごとに上位 10 件の宿泊施設に対する被験者評価の平均を示している．

5.3 考察

表 1 の実験結果から，全体の平均は 3.23 であり，被験者の評価はよく，主観的なクエリに対するシステムの有効性が確認された．

評価が低いクエリの原因は以下の通りである．クエリ (8) から要求情報 (宿, -, 落ち着く), (宿, 朝食, とれる) が抽出される．この場合,「朝食が取れる」,「宿が落ち着く」という意味から検索されるためクエリの意味をあまり考慮した検索ができなかった．クエリ (2) とクエリ (9) では，これらの要求情報に完全に一致する意見情報はほとんど存在せず，例えば，クエリ (9) の場合，(宿 A，部屋-値段, 高い) など，一致率のより低い意見情報が取得されてしまった．

6. おわりに

本論文では，自然言語で表現された検索クエリの主観性にロバストな商品検索システムを提案した．口コミ情報を活用した商品検索により，検索クエリにおける言語表現の多様性に対応することができる．宿泊施設を対象とした商品検索システムを作成した．約 22 万件の意見テキストを使用して検索実験を実施した結果，本システムの有用性を確認した．

今後は，意見情報と要求情報との対応付けにおいて，語彙の多様性に対処することが必要である．シソーラスを活用した自然言語による検索方式について検討することを予定している．

参考文献

[1] M. Dittenbach, D. Merkl, and H. Berger. A Natural Language Query Interface for Tourism Information. *Proceedings of the 10th International Conference on Information Technologies in Tourism (ENTER 2003)*, pp. 152–162, 2003.

[2] M. Hu and B. Liu. Mining Opinion Features in Customer Reviews. *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI 2004)*, pp. 755–760, 2004.

[3] B. Liu, M. Hu, and J. Cheng. Opinion Observer: Analyzing and Comparing Opinions on the web. *Proceedings of the 14th International World Wide Web Conference (WWW 2005)*, pp. 342–351, 2005.

[4] 小林, 乾, 松本, 立石, 福島. テキストマイニングによる評価表現の収集. 情処研報, NL-154, pp. 77–84, 2003.

[5] 立石, 福島, 小林, 高橋, 藤田, 乾, 松本. 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. 情処研報, NL-163, pp. 1–8, 2004.

[6] 黒橋. 日本語構文解析システム KNP version 2.0, 2005.