カスタマーレビューに基づく商品検索システム

杉 木 健 二[†1]　松 原 茂 樹[†2]

インターネット利用者の増大に伴い，電子商取引 (EC) サイトが急増している．一般にこれらのサイトでは，膨大な商品の中からユーザの要求に一致する商品を探すために商品検索システムが提供されている．しかし、既存の商品検索システムは検索項目があらかじめ定めたものに限られており，ユーザの要求の多様性，主観性に対応できているとは言えない．本稿では，カスタマーレビューに基づく検索システムを提案する．また，レビューから感性シソーラスを自動構築する手法について述べる．商品に関するレビューと，レビューから自動的に構築したシソーラスを用いることにより，ユーザが入力する自然言語文における表現の多様性，主観性に対応できる．最後に，本システムのプロトタイプである，宿泊施設検索システム「宿探」について述べる．

Product Retrieval System based on Consumer Evaluation

Kenji Sugiki[†1] and Shigeki Matsubara[†2]

In this paper, we propose a product retrieval system. Using consumer product reviews and the thesaurus developed from the reviews, our system can respond to a wide variety of users' requests, especially to subjective requests, to which existing systems can not respond. Finally, we describe an accommodation retrieval system named "Yado-tan" which is an experimental prototype of our system.

1. はじめに

近年，インターネット利用者の増大に伴い，楽天やAmazonをはじめとして，電子商取引 (EC) サイトが急増している．一般に，これらのサイトで扱う商品の数は膨大であるため，ユーザが目的とする商品に効率的にアクセスできる環境を提供することが望ましい．

これらのサイトでは一般的に，商品検索システムが提供されており，商品名，型番，または，メーカ，商品の性能，価格帯などによる商品の検索が可能である．しかし，それらのシステムでは，使用可能な検索項目はあらかじめ定められたものに限られており，ユーザが目的とする商品を検索できない場合が存在する．例えば，ユーザが自身の要求を具体化・数値化できない，また，要求に該当する検索項目が分からない場合がある．

これらの問題に対して，検索要求をユーザが自然言語で記述し，それを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[1)]，対話的な商品推薦システム[2),3)]などが提案されている．自然言語入力方式では，システム側が自然言語の表現の多様性，ユーザの主観性に対応する必要がある．しかし，クエリをデータベース言語に変換する方式では，変換ルールを人手で作る必要がありルールの網羅性に欠ける，または，変換対象となるデータベース項目が存在しない等の問題が生じる．

そこで本稿では，カスタマーレビューを用いた，消費者の評価に基づく商品検索システムについて述べる．また，ユーザの要求に関連した意見，及び，反対の意見を考慮するために利用するシソーラスの自動構築手法について説明する．最後に，本システムのプロトタイプである，宿泊施設検索システム「宿探」について述べる．

2. 意見に基づく商品検索モデル

多様な検索要求，または，主観性の高い検索要求に応えるために，本手法では，自然言語で表現された検索クエリに対して，要求に合致する情報が意見に記載されている商品を提示する．例えば，図1左の検索クエリに対して，商品「液晶テレビA」のレビュー中に図1右の意見文が存在すれば，「液晶テレビA」はこ

†1 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University
†2 名古屋大学情報連携基盤センター
Information Technology Center, Nagoya University

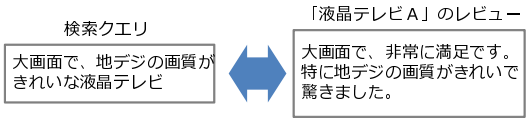

図 1　検索クエリに対応するレビューテキストの例

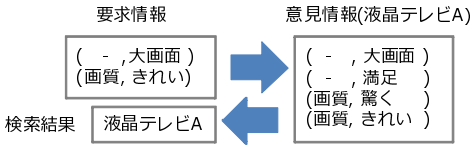

図 2　要求情報と意見情報の適合の例

のクエリに適合した商品であると見なす．検索クエリに対するレビューの適合性を測るために，クエリとレビューから (**項目**, **値**) の組をそれぞれ抽出する．ここで,「項目」は商品の属性や特徴を,「値」は商品の属性値や項目に対する消費者の評価を表す．これは，ユーザが商品を検索する場合,「色は赤でデザインがシンプルで音質がクリアな携帯プレイヤー」のように，商品の特徴とその値の組を検索条件とすることが多いためである．

レビューから抽出した各組を**意見情報**と呼び，一方，クエリから抽出した各組を**要求情報**と呼ぶ．意見情報と要求情報を比較し，項目及び値が等しければ，この意見情報は要求情報に適合すると判断できる．図 1 における例を図 2 に示す．

直接一致する表現だけでなく，類似もしくは反対の表現も考慮することができれば，検索再現性が向上し，よりユーザの要求に合致する商品を提示できると考えられる．本研究ではさらに，意見から自動的にシソーラスを構築することを試みる．

3. 商品検索システム

本システムの構成を図 3 に示す．本システムは，(1) 意見から意見情報の抽出，(2) 抽出した意見情報からシソーラス構築，(3) 意見情報とシソーラスを用いた商品検索の 3 つの処理部から構成される．以降，各処理について説明する．

3.1 意見情報の抽出

意見情報を構成する (項目，値) の組は，カスタマーレビューにおいて，(1) 主語-述語の関係（例えば,「音質が-クリア」），もしくは，(2) 修飾-被修飾の関係（例えば,「クリアな-音質」）として出現すると考えられる．文節間の係り受け関係としてみると，(1) は,「名詞＋(は/が/も)」のパタンを含む文節（項目）が用言の文節（値）に係る関係，一方，(2) は，連体形の用言を

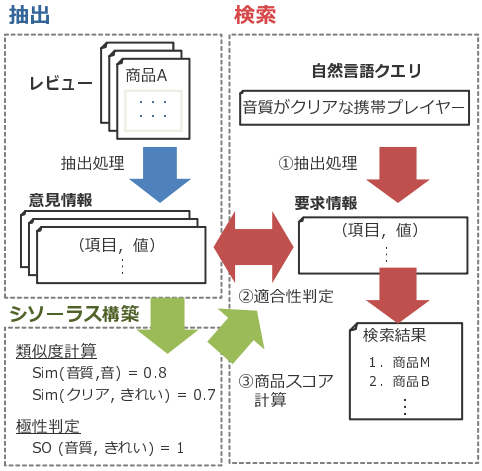

図 3　システムの構成

含む文節が名詞を含む文節に係る関係に対応する．これらの文節間の係り受けパタンを作成し，(項目, 値) の組を抽出する．ただし抽出の際，機能語などは除去する．本研究では，さまざまなドメインに容易に拡張できるように，ドメイン固有の辞書を用いず単純で自動的に抽出可能な方法を用いている．詳細については，文献[4] を参照されたい．

3.2 感性シソーラスの自動構築

要求情報の項目，値の文字列が一致する意見情報を扱うのみでは，要求情報と類似した意味や反対の意味を含む意見情報を捉えられない．そのため，検索再現性が低くなる，または，検索結果に適切ではない商品を含めてしまう可能性がある．

一般的に情報検索では，関連した概念を捉えるためにシソーラスによるクエリ拡張が行われる．シソーラスとは概念間の関係を示した語彙集であり，階層シソーラス (hierarchical thesaurus) と類似シソーラス (similarity thesaurus) に大別される．前者は，語彙間の上位・下位関係，部分・全体関係，同義・類義語などから成り，階層的な関係を定義している．一方，後者は，同義語・類義語の関係を扱い，語をノード，各ノード間の関連度をエッジとする重み付きの有向グラフのネットワーク構造をもつ．本研究では，後者の類似シソーラスを拡張した感性シソーラスの自動構築を試みる．具体的には，(項目, 値) の組をノード，組間の類似度をエッジとし，類似度が取りうる値を $[-1, 1]$ に拡張する．負の値の類似度を許可することにより，ある組と類似した概念を含む組，近い関係であるが意味的に反対の概念を含む組を区別して捉える事が出来る．

一般的に，シソーラスの自動構築には，2 単語間の文脈類似性を利用する[5]．これまで，周辺語や依存関

係などを文脈情報として利用する試みが多くなされてきた[6)–8)]．本研究では，２つの語彙が組み合わさった組間の類似度を求める必要があるため，２つの組の項目と値の類似度をそれぞれ独立で計算し，これらの積を組間の類似度とする．項目は値を要素ベクトル，一方，値は項目を要素ベクトルとし，類似度を算出する．これは，文脈情報として，主語-述語関係，修飾-被修飾の関係の依存関係を利用しているといえる．類似度の尺度には，コサイン尺度と２つのベクトル数が大きいほど値が大きくなる重みの積を与える．

以下に，２つのベクトル V_1, V_2 間の類似度の計算式を示す．

$$Sim(V_1, V_2) = g(V_1, V_2) \cdot cos(V_1, V_2)$$

$$g(V_1, V_2) = \log_2(\frac{|V_1| + |V_2|}{\max\_vec \cdot 2} + 1)$$

ただし，

$$Sim(V_1, V_2) = 0 \text{ if } Sim(V_1, V_2) < \alpha$$

ここで，max_vec は最大ベクトル数を示し，$g(V_1, V_2)$ はベクトル数の重みである．類似度が閾値 α 未満の場合 0 とする．このままでは，２つの組が反対の意味を示す場合も類似度が近くなってしまう（例えば，(部屋, 広い) と (部屋, 狭い)）．本研究では，各組の極性を導入する．極性とは，その表現が肯定的か否定的かを表し，肯定であれば極性の値を 1，否定的であれば値を-1 とする．以上から，組間の類似度は，項目間と値間の類似度の積に，さらに，各組の極性を掛けることにより算出される．

以下に２つの組 P_a ，P_b 間の類似度 $Sim(P_a, P_b)$ の計算式を示す．

$$Sim(P_a, P_b)$$
$$= Sim_f(f_a, f_b) \cdot Sim_v(v_a, v_b) \cdot SO(P_a) \cdot SO(P_b)$$

ここで，$Sim_f(f_a, f_b)$ は項目 f_a, f_b 間の関連度を示し，$Sim_v(v_a, v_b)$ は値 v_a, v_b 間の関連度を示す．$SO(P)$ は極性を表し，組 P が肯定であれば 1，否定であれば-1 を返す．要求と意見が同一の極性であれば類似度の値は正，異なる極性であれば類似度の値は負となる．図 4 に，感性シソーラスの一部を示す．感性シソーラスは，項目と値それぞれの，概念をノード，概念間の類似度をエッジとするシソーラスと，(項目, 値) 組の極性リストに基づき構築される．

極性判定には，那須川らの「評価表現の文脈一貫性」[9)] を利用する．文脈一貫性とは，ある対象に関する評価を記述する際，好評または不評の極性の意見を列挙することが多く，極性が反転する際には接続表現で明示することが多いという経験則である．文脈一貫

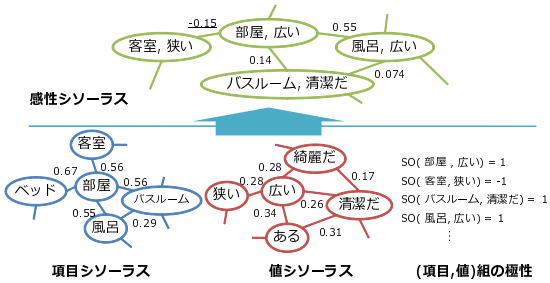

図 4　感性シソーラスの一部

性を利用し,「満足する」「不満だ」などの種表現の周辺文脈から評価表現の候補とその極性を抽出する．各候補の文書全体における分布から評価表現としての妥当性を判定する．得られた評価表現を種表現に追加し、これらの操作を再帰的に繰り返す．ただし本研究では，抽出処理を別に設けており，抽出した意見情報に対し極性を付与するのみとする．

3.3　カスタマーレビューに基づく商品検索

3.3.1　要求情報の抽出

3.1 節と同様の手法により，自然言語で記された検索クエリから要求情報の集合を抽出する．本研究では，検索条件及び要求対象から構成される名詞句によってクエリを表現することを想定しており（例えば,「音質がクリアな携帯プレイヤー」），名詞句の主要語 (head) を要求対象として抽出する．例えば，検索クエリ「音質がクリアでサイズが小さい携帯プレイヤー」からは，要求情報集合 { (音質, クリア), (サイズ, 小さい) } を抽出する．

3.3.2　商品のスコアリング

抽出した要求情報と意見情報から，ある商品のスコアを計算する．商品スコアを計算するために，要求と関連した意見，かつ，要求と反対の意見がどの程度含まれるかを考慮する．つまり，関連した意見を含めた意見の中で要求と同じ意見の割合がより多く，かつ，その意見の頻度が高ければ，その商品は信頼できる商品であると見なす．これらの要求と関連した意見，反対意見を感性シソーラスから算出される類似度に基づいて商品スコアに反映させる．

ある要求 Q における商品 P のスコアの計算式を以下に示す．

$$Score(Q,P) = \sum_{q_i \in Q} l_i g_i$$

$$l_i = \frac{\sum_{p_j in P} Sim(q_i,p_j) \cdot freq(p_j)}{\sum_{p_j in P} |Sim(q_i,p_j)| \cdot freq(p_j)}$$

$$g_i = \log\left(\frac{\sum_{p_j \in P} S\_Sim(q_i,p_j) \cdot freq(p_j)}{\sum_{p_j \in P} S\_Sim(q_i,p_j) \cdot avg\_freq(p_j)}\right)$$

ここで，要求 Q は要求情報 q の集合であり，商品 P は意見情報 p の集合である．$Sim(q_i,p_j)$ は感性シソーラスから求めた類似度であり，$avg\_freq(p_j)$ は組 p_j の意見全体における平均頻度である．l_i は，ある商品における要求と同一の極性の意見が含まれる割合を示す式である．g_i は，組 q_i の意見全体における相対的な出現頻度を示す式であり，各組の平均との比を求めることにより各意見間の頻度のばらつきを抑えている．

4. 宿泊施設検索システム「宿探」

本システムのプロトタイプとして，宿泊施設を対象とした検索システム「宿探」を開発した．利用者は，宿泊施設の検索クエリを自然言語で自由に記すことができる．カスタマーレビューとして，宿泊予約サイト「楽天トラベル」の意見投稿サイト「お客さまの声」[⋆1]を使用した．20,588 施設に対する意見テキスト 700,826 件を登録している．係り受け解析には KNP[10] を使用した．また，楽天 Web サービス[⋆2]を利用することにより，ホテル情報の提示や地域の絞り込み，空室検索に対応している．名古屋市内の「料理がおいしくて，部屋が広い宿」の検索結果を図 5 に示す．検索結果では，ホテルの情報に加え，各宿泊施設に関する意見のタグクラウドや，クエリと関連のある口コミ情報を提示することが可能である．

5. おわりに

本稿では，カスタマーレビューに基づく検索システムを提案した．また，カスタマーレビューから感性シソーラスの自動構築する手法を提案した．商品について書かれた意見テキストを使用することにより，ユーザが入力する自然言語文における表現の多様性，主観性に対応可能である．本システムでは，意見から意見情報の抽出，感性シソーラスの構築が自動で行えるため，さまざまなドメインへの拡張が容易である．今後は，システム性能の定量的評価，複数の種類の商品を対象としたシステムの構築に取り組む予定である．

⋆1 http://travel.rakuten.co.jp/auto/tabimado_bbs_top.html
⋆2 http://webservice.rakuten.co.jp/

図 5　「宿探」の検索結果

参　考　文　献

1) Dittenbach et., al.: A natural language query interface for tourism information, *ENTER-2003*, pp.152-162 (2003).
2) Chai et., al.: Natural language assistant: a dialog system for online product recommendation, *AI Magazine*, Vol. 23, No. 2(2002).
3) Mcsherry, D.: Explanation in recommender systems, *Artificial Intelligence Review*, Vol. 24, No. 2(2005).
4) 杉木健二, 松原茂樹: 消費者の意見に基づく商品検索, 情報処理学会論文誌, Vol. 49, No. 7, pp.2598-2603 (2008).
5) Harris, Z.: Distributional structure, *The Philosophy of Linguistics*, Oxford University Press, pp. 26-47 (1985).
6) Ruge, G.: Automatic detection of thesaurus relations for information retrieval applications, *Foundations of Computer Science*, LNCS, Vol. 1337 pp. 499-506, Springer Velag (1997).
7) Lin, D.: Automatic retrieval and clustering similar words, *In Proceedings of COLING/ACL 1998*, pp. 786-774 (1998).
8) Lowe, W. and McDonald., S.: The direct route: mediated priming in semantic space, *In Proceedings of the 22nd Annual Conference of the Cognitive Science Society (CogSci '00)*, pp. 675-680 (2000).
9) 那須川哲哉, 金山 博: 文脈一貫性を利用した極性付評価表現の語彙獲得, 情報処理学会研究報告, Vol.2004, No.73, SIG-NL-162, pp. 109-116 (2004).
10) 黒橋禎夫：日本語構文解析システム KNP version 2.0 (1998).