カスタマーレビューに基づく商品検索のための感性表現シソーラスの構築

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

{sugiki, matubara}@nagoya-u.jp

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．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，事前に設定された検索項目に限定されているため，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である．ユーザが目的とする商品を検索可能とするためには，ユーザの検索要求の多様性，または，主観性に対応する必要がある．

これらの問題を解決する方法として，ユーザの検索要求を自然言語により表現し，商品検索システムへの入力とする方法が考えられる．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]などが提案されてきた．しかし，クエリをデータベース言語に変換する方式では，変換ルールの網羅性に欠ける，あるいは，変換対象となる項目が存在しない等の問題が生じる．

そこで我々は，これまで，商品レビューに基づいた商品検索方式の開発に取り組んできた[4]．自然言語表現による検索クエリを入力とし，要求と一致する表現がレビュー上に存在すれば，そのレビューが意見の対象としている商品を提示する．商品レビューには，消費者の視点に基づく情報を含んでいるため，大量の商品レビューを用いることにより，ユーザの多様な要求に対応が可能である．特に，これまでほとんど不可能であった主観的な要求への対応が可能となった．しかし，表層的な表現の一致に基づき検索を実行していたため，類似，または，反対の意味を含む意見を含めることができず，検索の再現性，ランキングの適切さに問題があった．この問題を解決するため，本研究では，商品レビューから感性表現シソーラスを構築し，このシソーラスを用いた商品検索方式を提案する．評価実験により、シソーラスを用いた場合の検索性能について検討した．

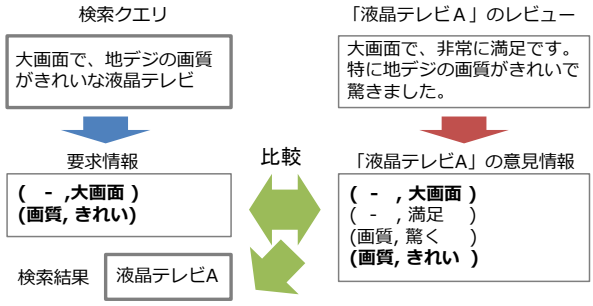

図 1: 本手法における検索方式

2 意見に基づく商品検索方式

本手法における商品検索方式について概説する．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れば，その意見対象の商品を提示する．ユーザの入力として，図1左上に示すような自然言語表現を想定する．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．この検索クエリに対して，図1右上に示すような「液晶テレビA」のレビューが存在する場合，検索クエリと内容が一致しているので，「液晶テレビA」はこ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それぞれ(**項目**, **値**)の組を抽出する．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．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の場合，(-, 大画面)，(画質, きれい)の組が一致しているので，「液晶テレビA」を検索結果として返す．

検索の再現性を高め，ユーザの要求により合致する商品のスコアを高くするためには，要求と意見の組が完全に一致している組だけでなく，類似もしくは反対の意味を含む組も考慮する必要がある．これらの組間の類似度を求めるための感性表現シソーラスの構築と，このシソーラスを用いた商品検索方式について以下に説明

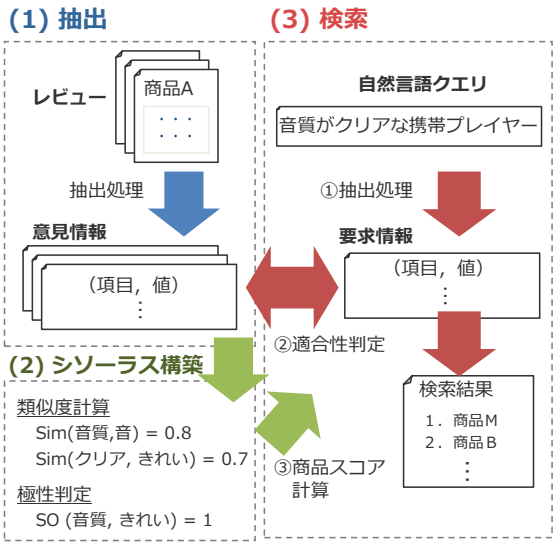

図 2: システムの構成

する．

3 商品検索システム

本システムの構成を図 2 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

商品レビューにおいて，意見情報を構成する (項目, 値) の組の多くは，(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，(2) は，連体形の用言を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，用言のみ存在する場合は，(-, 値) として組を抽出する．抽出の際，機能語などは除去する．また，ノイズ除去のため，組の全体における出現頻度が閾値以下となる場合はフィルタリングする．

3.2 感性表現シソーラスの自動構築

本研究では，２つの語彙が組み合わさった組間の類似度，さらには組間の関係が反義の表現を含む場合も表現可能なシソーラスを構築する．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，このうち，類似シソーラスを拡張した**感性表現シソーラス**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の範囲を $[-1, 1]$ に拡張する．

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, 7, 8]．一方，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に計算し，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，値は項目を要素ベクトルとして，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．類似度の尺度として，Tanimoto 係数 [9] (extended Jaccard coefficient) を用いた．以下に，２つのベクトル V_a, V_b 間の類似度 $T(V_a, V_b)$ の計算式を示す．

$$T(V_a, V_b) = \frac{V_a \cdot V_b}{||V_a|| + ||V_b|| - V_a \cdot V_b}$$

ただし，$T(V_a, V_b)$ が閾値 α 未満の場合，$T(V_a, V_b) = 0$ とする．

さらに本研究では，組間の類似度に負の値を導入するため，各組の極性を判定する．極性とは，その表現が肯定的か否定的かを表し，肯定的であれば 1，否定的であれば-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．極性判定には，那須川らの「評価表現の文脈一貫性」[10] 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．本研究では，極性された組の頻度が β 以上，かつ，その組に対して同一の極性の割合が閾値 γ 以上の場合にその極性を付与する．さらに，すべての組に極性を付与するため，組の極性が判定されなければ，その組に含まれる値の極性を適用する．さらに，値の極性が存在しない場合はその組に対して肯定を付与する．否定表現が含まれる場合は極性を反転させる．

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．

$$\begin{aligned} &Sim(p_a, p_b) \\ &= T(f_a, f_b) \cdot T(v_a, v_b) \cdot SO(p_a) \cdot SO(p_b) \end{aligned}$$

ここで，$T(f_a, f_b)$ は項目 f_a, f_b 間，$T(v_a, v_b)$ は値 v_a, v_b 間の類似度を示す．$SO(p)$ は極性を表し，組 p が肯定的であれば 1, 否定的であれば-1 を返す．要求情報と意

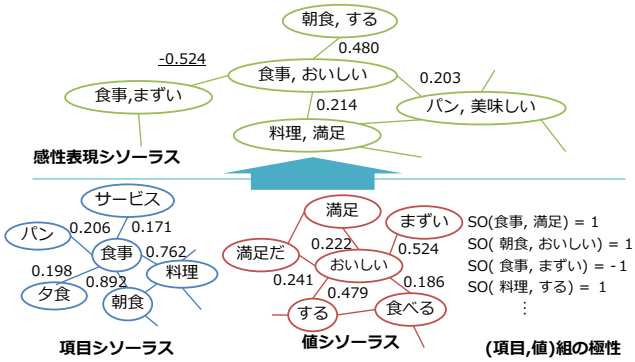

図 3: 感性表現シソーラスの一部

見情報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 3 に，感性表現シソーラスの一部を示す．感性表現シソーラスは，項目と値それぞれの類似シソーラスと組の極性リストに基づいて構築される．

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，要求情報の集合 { (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

抽出した要求情報の集合と各商品の意見情報の集合から，その商品のスコアを算出する．要求と同一の極性の割合が高く，かつ，より出現頻度が高い商品から順に提示する．

ある要求 Q における商品 P のスコア $Score(Q,P)$ の計算式を以下に示す．

$$Score(Q,P) = \sum_{q_i \in Q} R_i \cdot F_i$$

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

$$F_i = \log\left(\sum_{p_j \in P} Sim(q_i,p_j) \cdot freq(p_j) + 1 \right)$$

ここで，R_i は要求情報 q_i が商品 P においてどれだけ極性が一致するかを表し，F_i は要求情報 R の商品 P における出現頻度に相当する．$freq(p_j)$ は商品 P における意見情報 p_j の出現頻度を表す．

4 宿泊施設を対象とした検索実験

4.1 実験方法

本方式による検索性能を評価するため、消費者のレーティング評価を用いた検索実験を行い，シソーラスを用いた場合と用いない場合とで比較をする．対象ドメインを宿泊施設に設定し,「楽天トラベル」の意見投稿サイト「お客さまの声」[1]に記載されているレビューとレーティング評価を使用した．実験はそのうち 1000 施設を対象とし，その施設の意見テキスト 215,458 件を用いた．レーティングは，宿泊施設の利用者が，その施設について「総合」「サービス」「設備・アメニティ」「立地」「部屋」「風呂」「食事」の観点から 5 段階評価し，サイト上には投稿人数，及び，平均化された 0.5 刻みの 9 段階での評価が記載されている.「総合」を除く 6 つの各観点において検索を実行し，システムとレーティングとの間のランキング順位の相関係数を求めることにより，シソーラスを用いた場合の有効性を評価する．ベースラインとして，商品スコアの計算式において要求情報と意見情報が完全に一致する（類似度 Sim の値が 1 となる）場合を設定した．各観点における検索クエリを，表 1 に示す．各観点に対して頻出する検索クエリを設定した．係り受け解析には KNP[11] を使用し，組の出現頻度の閾値を 10 と設定し，それ未満の場合除去した．また，極性判定については，10 回試行し，極性判定された組の出現頻度が 5 以上，かつ，極性判定のうち 8 割以上が同一の極性と判定された場合，極性リストに追加した．

4.2 複数の観点における検索とレーティングとの比較結果

6 つの観点における検索結果を表 1 に示す.「立地」の場合を除くすべての観点において，シソーラスを用いたほうが検索ヒット数が増加し，かつ，レーティングとの相関も高くなっていることが分かる．この結果から，再現性が向上し，かつ，よりユーザの評価に近いランキングとなっていると考えられ，感性表現シソーラスを用いることの有効性を確認できた．ただし，今回設定した検索クエリは，ある観点における一部の要求を表現したものであり，レーティングとの相関関係を利用して本システムとベースラインとを相対的に比較することを目的としている．

ベースラインと比較し相関係数の差が大きければ，シソーラスを用いた効果があったといえる．そのような例として,「食事」の観点の場合について述べる.「食事がおいしい」というクエリから要求情報 (食事, おいしい) が抽出される．項目「食事」に対して,「朝食」「夕食」「料理」「パン」の項目が類似度が高いと判定された．一方，値「おいしい」に対して,「美味しい」「する」「満足」「食べる」の値が類似度が高いと判定された．結果として，この要求情報に対して，(朝食, 美味しい), (食事, する), (パン, 美味しい), (料理, 満足) などの意見情報を含む宿泊施設を適切に検索可能であった．

[1] http://travel.rakuten.co.jp/auto/tabimado_bbs_top.html

表 1: 各観点における検索クエリと実験結果

観点	検索クエリ	相関係数 (検索ヒット数)	
		ベースライン	提案システム
サービス	「サービスが良い」	0.099 (271)	0.267 (974)
立地	「立地が良い」	0.154 (973)	0.154 (973)
部屋	「部屋がきれい」	0.183 (594)	0.352 (956)
設備・アメニティ	「設備が充実し、アメニティが充実している」	0.219 (209)	0.252 (956)
風呂	「風呂が広い」	0.203 (260)	0.313 (986)
食事	「食事がおいしい」	0.183 (125)	0.331 (781)

一方，相関係数の差が小さい場合，シソーラスが必ずしも適切に利用されなかった可能性がある．そのような場合として,「立地」の観点の場合について述べる.「立地が良い」というクエリから要求情報 (立地, 良い) が抽出される．項目「立地」に対して，関連した項目「立地条件」「場所」が類似度が高いとして判定されたが，同様に，あまり関係の無い項目である「サービス」「感じ」「対応」も，類似度が高いとして判定された．その結果，不適切な意見情報が含まれ，相関係数が小さいままとなった．

5 おわりに

カスタマーレビューに基づく商品検索のための感性表現シソーラスの構築方法と，シソーラスを用いた商品検索方式について述べた．評価実験では，感性表現シソーラスを用いることにより，検索再現性が向上し，検索性能がよりユーザの評価に近くなることを確認した．

今後は，複数の被験者による検索評価実験に取り組む予定である．また，今回獲得した意見情報のほとんどの極性が 1 と判定され，極性を用いることの有効性については確認できなかったため，感性表現シソーラスの評価をより詳しく行う．

参考文献

[1] Dittenbach et., al.: A natural language query interface for tourism information, In *Proceedings of 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 (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.: Lowe and McDonald: The direct route: mediated priming in semantic space, In *Proceedings of CogSci 2000*, pp. 675-680 (2000).

[9] Tanimoto, T.: IBM Internal Report 17th Nov. (1957).

[10] 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).

[11] 黒橋禎夫：日本語構文解析システム KNP version 2.0 (1998).