

インターネットからの評判情報検索

4W-5

立石健二 石黒義英 福島俊一
NECインターネットシステム研究所

1. はじめに

インターネットは、あらゆる個人が独自の情報を発信できる場である。従って、そこには他のメディアにはないさまざまな人々の「生の声」が存在する。このような情報は、商品購入やマーケティングの際に重要な意味を持つが、従来のキーワード型検索では効率的に情報を収集することが難しかった。

今回筆者らは、インターネットから任意の商品に関する評判情報を自動検索する新しいサーチエンジンを提案する。これは、筆者らが進めていた目的特化型サーチエンジン[1]のアプローチに添うものである。また、従来の商品検索エンジンである ShopBot[2]等が、価格やスペックのような定型情報を検索対象としているのに対して、本エンジンは、評判情報という不定形な情報を対象とすることが特徴である。

2. 評判情報検索機能

評判情報検索は、あらかじめクローラがインターネットから集めてきた HTML ページ群の中から、ユーザが入力した商品に関する評判情報を含むページを検索し、ユーザに提示するものである。

検索結果は、入力された商品名に関する評判情報を含む HTML ページのタイトル、URL、ページ内容を一覧で表示するが、ユーザが商品の良し悪しを容易に理解できるよう次のように工夫した。

- (1)優先順位判定機能—文脈的に評判情報らしいものほど結果一覧の上位に位置するようにする。
- (2)分類機能—評判情報の内容が肯定的か否定的か、又はその評判を誰が記述しているかを分類して示す。
- (3)ピンポイント表示機能—ページの先頭ではなく、評判情報が出現する部分を切り出して表示する。

3. 検索方式

前節の機能を実現するにあたり、評判情報とは例えば「モバイルギアは良い」のように、検索語と評価表現を含む一定範囲の文字列と定義した。

評判情報検索では、ユーザは検索語と共に商品カテゴリを入力する。検索手順は次の図 1 のようになる。

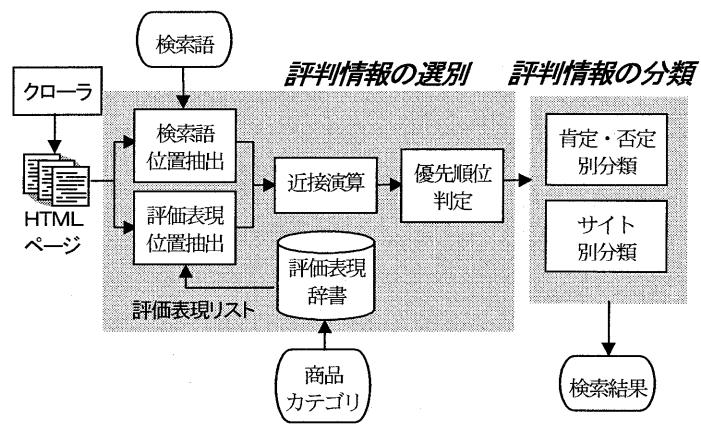


図 1 検索手順

3. 1 評価表現辞書

評価表現とは、物事に対する人間の主観的な判断を示す表現とした。品詞に着目すると「良い」「好き」「最高だ」「面白い」といった用言に多いが、それ以外にも例えば「人気」「絶品」「駄作」といった名詞も該当する。また、全体的に人間の感情を示す表現が多いが、「速い」「重い」等の物の性質を示す表現が使われることもある。

評価表現は商品カテゴリ毎に作成し、あらかじめ辞書として用意しておく。商品カテゴリとは、例えば「書籍」「コンピュータ」「車」である。これは、評価表現には、「いい」「悪い」等のすべてのカテゴリに共通の表現だけでなく、「旨い」「読みやすい」等、カテゴリ固有の表現が存在するためである。

3. 2 近接演算

入力された検索語と評価表現を近接演算して、マッチする場合は、検索語と評価表現を含む一定範囲の文字列（評判情報）を抽出する。評価表現は、評価表現辞書の中で入力された商品カテゴリに適合する評価表

現リストのすべてである。そのため、具体的には次の検索式の演算を行う。

$W_p \text{ NEAR } (W_{s1} \text{ OR } W_{s2} \text{ OR } \dots \text{ OR } W_{sn})$

W_p 検索語 $\{W_s\}$ 評価表現リスト

3. 3 優先順位の決定

優先順位の決定は、まず近接演算処理で抽出した評判情報に対して正規表現で記述された表1のような優先度判定ルールを適用する。

ID	優先度判定ルール
1	$_KEY_*(\text{は} \text{が} \text{も})*_REP_$ (検索語と評価表現が「は」「が」「も」を挟んで存在する)
2	$_KEY_*(. ? !)*_REP_REP_*(. ?$ $!)*_KEY_$ (検索語と評価表現が別の文にある)

表1 優先度判定ルールの例（ $_KEY_$ は検索語、 $_REP_$ は評価表現を示す）

次に、総ルール数が n のとき n 次元の配列を用意し、 n 個のルールを満たすか満たさないかを $1/0$ の値でセットする。例えば、総ルール数が 4 つで評判情報が ID1 と ID3 のルールを満たす場合は(1,0,1,0)である。各配列のパターンにはスコアを対応付けてあり、このスコアが評判情報の優先度を示す。なお、各配列パターンのスコアはあらかじめ用意したサンプル評判情報に対して人間が付けたスコアを配列パターン毎に平均値を取ることにより決定した。

3. 4 評判情報の分類

評判情報の分類は、肯定・否定別の分類とサイト別の分類がある。肯定・否定別の分類は、評価表現の近傍に否定表現が出現したか否かで判定する。否定表現とは、例えば「ない」「ません」「なかった」「不（接頭語）」のような、評価表現の意味を打ち消す役割をもつ表現である。サイト別の分類は、評判情報を含む HTML ページの URL を用いて、あらかじめ決められたサイト種に分類する。サイト種は、評判情報が良く出現するサイトの種類であって、例えば「ニュースサイト」「掲示板」「ショッピングサイト」が該当する。

4. 試作システム

図2に3節の検索方式に従い作成した評判情報検索システムの検索語「モバイルギア」商品カテゴリ「コンピュータ」に対する検索結果を示す。用意した評価

表現の数（コンピュータ関連）は 100 程度である。また、近接演算の距離は、検索語を基準として前方 40Byte、後方 80Byte とした。各ページの評判情報は優先順に並んでおり、そのスコアは文脈的に評判情報らしい程高いスコアを与え最高 5 点としている。また、分類結果として肯定的又は否定的かを示すアイコン、サイト別のアイコンをそれぞれ表示している。

この試作システムの適合率（指定した商品名に関する評判が検索される精度）を測定したところ、上位 10 件で 90%、上位 20 件で 75% であった（16 個の検索語の平均値）。なお、単純に検索語を含むページ数は 16 個の検索語の合計で 9390、その内評判情報を含むページに限定することで 12% に絞り込まれた。



図2 検索結果画面

5. おわりに

本稿では、評判情報検索の機能及びそれを実現するための検索方式について述べた。本検索方式は、オンラインショッピングにおける商品購入支援や、インターネット上の情報モニタリングなどへの応用が考えられる。今後の課題は検索の高速化、評判の重要性に基づいた優先度判定等である。

参考文献

- [1] 福島俊一 石黒義英 喜田弘司 山田洋志 松田勝志, "目的特化型サーチエンジンの開発", 情報処理学会第 61 回全国大会予稿集(3), p137-138, 2000
- [2] Doorenbos, R., Etzioni, O., and Weld, D. "A scalable comparison-shopping agent for the World-Wide Web.", In Proceedings of the First International Conference on Autonomous Agents Agents'97(Marina del Rey, Calif., Feb. 5-8) ACM Press, N.Y., 1997, pp.39-48.