レビューページ例からの属性抽出に基づく レビューページ検索

赤木　法生† 　大島　裕明†† 　小山　聡†† 　田島　敬史†† 　田中　克己††

† 京都大学工学部情報学科
〒 606-8501 京都市左京区吉田本町
†† 京都大学大学院情報学研究科 社会情報学専攻
〒 606-8501 京都市左京区吉田本町
E-mail: †{akagi,ohshima,oyama,tajima,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　近年，Web 上の情報量は飛躍的な速度で増加し，利用者が，膨大な情報の中から求める情報を効率的に見つける事が難しくなっている．そうした情報の 1 つとして，ある商品のような対象に関する，評価や評判の情報に焦点を当てる．こうした情報は，対象についてどのような観点から書かれているかという点が重要であり，またどのページもそれらの観点に基づいて述べられているという傾向がある．したがって，各ページの内容を考慮してランキングされているわけでは無い検索エンジンよりも，そのような評価の対象となるポイントに基づいた情報検索が有効であると考えられる．本研究では，ある対象についての評判情報を記載したページ集合を一定数集め，それらから対象についてどのような評価のポイントがあるのかを抽出し，それらのポイントについて各ページでどの程度言及されているのかを尺度化する．こうする事で，ユーザが，自分の見たいページや，既に見たページには無いような情報が載っているページを検索する事ができるようにする．
キーワード　情報検索，知識発見，Web とインターネット，評判情報

Review Page Retrieval by Attribute Extraction from Review Page Samples

Norio AKAGI†, Hiroaki OHSHIMA††, Satoshi OYAMA††, Keishi TAJIMA††, and Katsumi TANAKA††

† Informatics of the faculty of Engineering, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
†† Graduate School of Infomatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
E-mail: †{akagi,ohshima,oyama,tajima,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　In recent years the amount of information on the Web is in rapid increase, and it becomes more difficult to find information people want efficiently. The Web contains a wealth of review information about various objects. In this paper, we focused on the way to get more review information efficiently. We gathered the example of the review pages about some object, and extracted features about the object by analyzing those pages. Then we examined how those pages are detailed about each feature, and people can see pages which contain new information that have not appeared in the preceding pages.
Key words　Information retrieval，Knowledge discovery，Web and the Internet，Reputation information

1. はじめに

現在では，インターネットというメディアを通して様々な情報が得られるようになっている．また Weblog の普及などによって，個人の主観的な意見を容易に発信出来るようになり，ユーザの主観に基づいて記述されたコンテンツも，Web 上で増

加してきている．本論文ではこうしたコンテンツの中で有用なものの 1 つである，ある対象 (特定の商品など) の評価や評判情報に焦点を当てる．

ある商品について，ユーザがその評価や評判に関する情報を調べたいと思ったとき，Amazon [1] や価格.com [2] のような，ユーザのレビューが載っているページを見たり，Google [3] などの検索エンジンで，適当なクエリを用いて検索を行い，検索結果として出てきたページを見るのが一般的である．しかし，例えば Amazon や価格.com のようなレビューサイトの存在を知らないユーザは，そのようなページに直接アクセスして情報を得る事ができない．また，検索エンジンを用いる際には，たとえばある商品に対してユーザが何か気になるポイントなどがあれば,「商品名＋そのポイント」のようなクエリを作ることができるが，商品に関する知識が無いユーザにとっては，このようなクエリを思い付くことは困難である．このような，ある商品に関して多くの人が評価の対象として言及しているポイント (例：デジタルカメラの場合，画素数や本体の大きさ，電池の持続時間など) を，本論文では商品の “評価属性語” と呼び，評価属性語に関する具体的な意見や評価を，“評価属性値” と呼ぶ．評価属性語は対象のジャンルごと，あるいは 1 つ 1 つの対象ごとに固有のものと考えられる．上記のような方法では，いずれの場合もユーザの対象に関する知識が前提となっているため，知識を持っていない人ほどより知識を得にくくなってしまっているという問題点がある．

また，他には検索エンジンを用いて「対象名」や「対象名 AND レビュー」などの簡単なクエリで検索を行うという方法も考えられるが，この場合，ユーザが多く観点からの情報を得たいと思えば，検索結果のページを順に 1 つ 1 つ見ていく必要がある．しかし，検索エンジンの検索結果は，各々の Web ページで対象のどんなポイントに関して述べられているかを考慮しているわけでは無いため，1 つページを新しく見るのに応じて．それまで見たページにはのっていなかった情報が確実に得られるというわけでは無い．さらに，検索結果のページには，あまり評価・評判情報がのっていないページも存在するため，この方法は効率的であるとは言えない．

本論文では，まずユーザが想定している商品などの対象について，その対象の評判や評価情報が載っているページをいくつか，Web ページ例として取得する．そして，それらの Web ページから対象の評価属性語を抽出し，各ページがそれぞれの評価属性語についてどの程度言及されているかを尺度化する手法を提案する．そのようにする事で，ユーザはページ例中のページが，対象の何についてどの程度語っているのかという特徴をとらえることができ，集合中から自分の読みたいページや，これまで読んだページとは内容が異なるようなページを探し出す事ができると考えられる.「内容が異なるページ」とはこの場合，

- 対象に関して，クエリとなるページでは言及されていないような評価属性語があった場合，それについて言及されているようなページ
- クエリとなるページで言及されている評価属性語について，より詳しく言及されているページ

の事を指す．このようにページを各評価属性語の言及度に基づいて数値化する事で，ユーザは，ある対象に関する評価・評判情報について，対象の持つ様々な評価属性語に関する，否定・肯定に関わらない多様な情報を，より多く，より詳細に，効率よく得る事ができると考えられる..

2. 基本的事項及び関連研究

2.1 基本的事項

2.1.1 商品の評価・評判情報に関する Web ページ

Web 上において，商品などの対象に関する評価・評判情報が載っているページの数は非常に多い．たとえば商品で言えば，実際にその商品を販売している企業の，製品情報のページも一つの評価・評判情報の載っているページと考える事もできるが，本研究においては，オフィシャルサイトなどの情報よりもむしろ，例えばデジタル電化製品などの情報 Web サイトにあるような商品の解説ページやレビューページ，あるいは実際に商品を買った個人の感想の載っているページなどを中心に考える事にする．商品以外の対象についても同様である．

個人ユーザの中には，有志で商品に関するまとめサイトを作ってレビューをしたり，Amazon.co.jp などのレビューが書き込める Web サイトにレビューを書いたりする人がいる．また，近年では Weblog や wiki の普及により，HTML 文書のソースを直接書くといった操作を通さずに手軽にコンテンツを作成することができるようになった．これにより，多くのユーザが個人で自分の感想を Web 上に公開する事ができるようになり，中には詳細なレビュー記事を書くユーザも存在し，まずます Web 上における商品などの評価・評判情報は増加していると思われる．

2.1.2 茶　　筌

茶筌 [4] は，奈良先端科学技術大学院大学の自然言語処理講座で開発されている，日本語の形態素解析エンジンであり，他のシステムから簡単に利用ができるようになっている．本研究では，最初に商品のレビューページ例を集めてそれを解析する事でその商品の評価属性語を抽出しているが，その際に茶筌を利用した形態素解析を行っている．また，各ページについてそれぞれの評価属性語についてどの程度言及化されているかを尺度化する際にも，評価属性語抽出の際の茶筅の解析結果を用いている．

2.2 関 連 研 究

2.2.1 Web 上からの評判情報の抽出に関する研究

Web 上からの評判情報の抽出の研究で最も多いのは，Web 上のページから評判文を抽出するというものである．

例えば，立石ら [5] の研究では，あらかじめ用意した評価表現辞書を基に，Web から意見情報を抽出し，否定・肯定の判定を試みている．また，鈴木ら [6] の研究では，収集された Blog のデータから評判情報らしき語彙の組を抽出して，教師データを基に非評価・肯定・否定の分類器を構築し，実験をしてその精度を評価している．他にも，藤村滋ら [7] の研究では，文を構成する主要な語のみを用いた，n-gram を素性を用いて，あらかじめ収集しておいた肯定・否定の評判から評価表現を抽出

している．

これらの研究では，いずれも Web 上の評価情報に対して，その中の評価・評判情報と思われる文章のみを抽出する事を目的とし，またその評価・評判情報が否定か肯定か，という部分に着目した抽出が行われている．本研究とは，評価・評判情報がその商品について否定的か肯定的かに関わらず，商品に関する幅広い観点からの評価・評判情報を効率よく得ることに着目しているという点で異なる．

2.2.2 感情マイニングに関する研究に関する研究

感情マイニングに関する研究としては，熊本らによる，Web 上のニュース記事から喜怒哀楽を抽出するという研究がある [8]．人間がニュース記事を読んだときに生じると思われる感情を「怒る」,「喜ぶ」と「悲しい」,「嬉しい」という四つに分類し，それらの単語と他の共起度を元に，各記事を読んだユーザがどのような感情（怒る⇔喜ぶ，悲しい⇔うれしい）を抱くかを尺度化して推定する手法を提案している．

本研究とは，感情語彙を対象として記事を読んだ際の感情を推定するのでは無く，商品の評価属性語を対象として，その記事でどれだけその評価属性語について言及されているかを推定するという点で異なるが，ベースとなる単語とその他の単語との共起を用いるアプローチを参考とする．

3. 評価属性語に注目した評判情報検索

3.1 「似て非なる検索」の概念

本研究の目的は，ユーザが，特定の商品などの対象の評価・評判情報が書かれているページの中で，自分が読みたいページを探したり，今まで見たページとは内容の異なるページを探し出してくる事である．ここで言う「内容の異なるページ」とは，前述のとおり，商品について，クエリとなるページには無い評価属性語について載っているページ，あるいはクエリとなるページにある評価属性語について，より詳細に載っているページと言う意味である．これは従来のような，ベクトル空間モデルにおける，tf-idf を用いた文書間類似度などでは求める事ができないため，別に尺度を考え，その尺度に基づいて文書間の類似度を計算する必要がある．このように，あるページに対して，類似の属性語集合を有しているが，その属性値が異なるような他のページや，元のページには含まれていない新たな属性語と属性値を含むような他のページを探し出すような検索を，本論文においては,「似て非なる検索」と呼ぶ。

本研究における「似て非なる検索」は,「同一の対象について，異なる評価属性語や評価属性値を持っているページを検索する」という部分にのみ言及している．対象の評価・評判情報の検索に関しては，その他にも,「同じ評価属性語や評価属性値を持つ，異なる対象についてのページを検索する」という「似て非なる検索」もあり，これは競合する別の商品の評判情報を検索するシステムなどへの応用が考えられる．本研究におけるアプローチは，こちらの「似て非なる検索」への拡張も期待できるものである．

3.2 提案・一連の流れ

最初に，ユーザが評価や評判情報を知りたいと思っている商品があるとして，その商品名は分かっているとする．

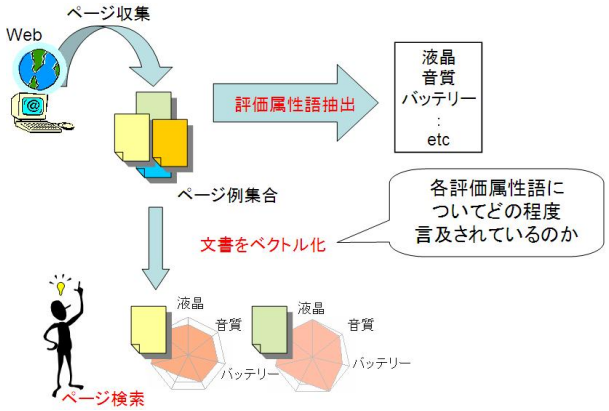

図 1 システムの概要

本提案手法において想定している一連のシナリオの流れは，以下の通りである．図 1 に，この流れを図示している．

（1） ユーザが，評価や評判情報を知りたい商品名を入力する．

（2） システムは，その商品に関する評価や評判情報が載っていると思われる，一定数のページ集合を用意する．

（3） システムは，ページ集合から，その商品について評価の対象となっているポイント (評価属性語) を抽出する．

（4） システムは，ページ集合の各ページについて，抽出した各評価属性語についてどの程度言及されているのかを尺度化し，各ページの文書をベクトル化する．

（5） ユーザが，その言及度の尺度を基にページ検索を行う．

ページ検索については，さまざまな手法が考えられる．図 1 のように，各文書の各評価属性語に対する言及度をレーダーグラフで視覚化してユーザに直接閲覧する文書を選択させたり，ユーザが 1 つページを選択・閲覧している際に，システムがページ集合中の他のページを推薦ページとして提示する方法も考えられる．本論文では，本手法で求めた「言及度」が実際のページの内容を正確に反映しているかどうかを調べるために，後者の場合を想定した．

4. 評価属性語の抽出

4.1 文書集合の取得

最初に，ユーザが商品名を入力すると，その商品の評価や評判情報を含むページ集合をシステムが収集する．本研究においては，Google の検索エンジンを用いて，ユーザが入力した商品名に応じて「商品名 AND (評価 OR レビュー)」というクエリで検索した，上位 100 件のページを集合として用いる事にする．

システムは，取得したページ集合を解析して，商品の評価属性語を抽出する．そのために，収集した Web ページ集合は html ファイルであるから，まずはタグを取り除いてテキストに変換する．検索結果上位 100 件のページ集合は，100 個の文書集合となる．

4.2 形容詞の頻度算出

取得した文書集合を用いて，商品の評価属性語を抽出する必要がある．本研究では，商品の評価属性語は名詞であるとし，

その評価属性語について書かれている意見・評価を表す語句を評価属性値と定義している．例えば,「このデジタルカメラは電池の保ちが良い」「この携帯電話のサイズは大きい」という文章ならば，前者であれば評価属性語は「電池」，評価属性値は「保ちが良い」，後者であれば評価属性語は「サイズ」，評価属性値は「大きい」となる．しかし，評価属性語が名詞であると言っても，単純に全文書から名詞の登場頻度を求めても,「商品名 AND (評価 OR レビュー)」というクエリで検索した検索結果のページでは,「評価」や「レビュー」をはじめとして，他の名詞も多く登場する．また，例えば上の例であれば,「この携帯電話は大きい」といったように，必ずしも評価属性語が無くても意味が通じる評判文も多い．このような理由から，名詞にだけ注目しても，評価属性語の登場頻度が必ずしも高くなるとは限らず，評価属性語と思われる語句をうまく抽出する事は困難である．

そこで，評価属性値，その中でも形容詞が，評価文の中にきわめて高い確率で現れる事に着目した．上の例でも,「良い」「大きい」といった形容詞はまず省略されることは無い．

したがって，まずは形容詞に注目し，文書集合から，登場回数が多い形容詞を求める．

4.3 形容詞の近傍にある名詞の登場頻度算出

評価やレビューといったクエリを加えて検索しているので，文書集合全体で頻度が上位にある形容詞は，評価属性語について形容している評価属性値であると考えられる．

「この携帯電話のサイズは大きい」の例では，評価属性値となる「大きい」の直前の名詞句が評価属性語となっている.「このデジタルカメラは電池の保ちが良い」の例のように，形容詞と，評価属性語となる「電池」の間に何か他の名詞が入っている事もあり，また全ての評価文がこの例のように「(商品名) の (評価属性語) は (評価属性値) である」という理想的な構造になっているとは限らないが，一般的に，評価属性値となる形容詞の近くに存在する名詞句が評価属性語である可能性は高いと考えられる．したがって，上で求めた登場回数の多い形容詞について，文の区切りなどは無視して，その形容詞の近傍にある語句の内，名詞句だけを取り出す．このようにして，形容詞に近い順から前後 m 個 (m は適当な自然数) の名詞を調べてその登場回数を求めれば，登場回数の多い名詞が評価属性語である事が期待できる．

5. 文書間の類似度計算

評価属性語の抽出が完了したら，次にその評価属性語で表される商品の属性に関して，各文書でどの程度言及されているかを尺度化する．評判文には,「この携帯電話の画面は大きい」のようにはっきりと「画面」という評価属性語が明示されている文と,「この携帯電話は大きい」のように，はっきりと書かれてはいないが,「サイズ」のような大きさを表す評価属性語に関して述べている文とがある．この例を見れば分かるように，評判文に必ずしも評価属性語が書かれているとは限らないため，システムが，1 つ 1 つの文章を解析して，ある文章がどの評価属性語について述べられているのかを判断するのは困難である．そこで本研究では，評価属性語との共起度を用いて，その文書がある評価属性語についてどの程度述べられているかを尺度化する手法を提案する．この,「ある評価属性語についてどの程度述べられているか」について，本研究ではこれをその評価属性語についての「言及度」と呼ぶことにする．

5.1 共起度を計算

収集した全文書中の名詞，形容詞，形容動詞，動詞 (非自立なものは除く) からなる単語の集合を W とする．W に含まれる各単語 w_i を対象に，評価属性語 t_j と 1 つの文章中で共起している度合いを計算する．共起度には Jaccard 係数を用いた．すなわち，集合 W 中のある単語 w_i と，共起度を求める対象となる評価属性語 t_j の間の共起度は，

$$\text{共起度 } Co_{ij} = \frac{w_i\text{を含み，かつ } t_j\text{を含む文章の数}}{w_i\text{または } t_j\text{を含む文章の数}}$$

で計算ができる．全文書中の各文章を解析して上の計算を行い，W の要素となる単語全てについて，評価属性語 1 つ 1 つに対する共起度を求める．

5.2 文書の評価属性ベクトル化

評価属性語が全部で q 個であるとすると，1 つの文書の，全ての評価属性語についての言及度は，q 次元のベクトル $V = (v_1, v_2, \cdots, v_q)$ として表すことができるはずである．ここで $v_j(1 \leqq j \leqq q)$ は，その文書における評価属性語 t_j についての言及度である．共起度が求まったら，今度は文書 1 つ 1 つに対して，$w_i \in W$ の出現回数 Ap_i を計算し，ある 1 つの文書の，全評価属性語に対する言及度を表すベクトル $(v_1, v_2, \cdots, v_q)$ において，

$$v_j(1 \leqq j \leqq q) = \sum_{w_i \in W} Co_{ij} \times Ap_i \qquad (1)$$

という計算で各要素のベクトルを求める．これを全文書に対して行うことで，全ての文書が評価属性語 t_j の言及度でベクトル化できる．

5.3 類似度計算

各文書が，各評価属性語の言及度によるベクトルで表されたので，文書間の類似度を，様々な方法で調べる事ができる．本研究では，提案手法により計算した言及度によるベクトルが，実際の文書における言及度をどれだけ正確に反映しているかを調べるために，ユークリッド距離を用いる方法を考える．すなわち，ユーザが検索エンジンの検索結果からあるページを選択したら，そのページと距離が最も遠く，かつ絶対値がそのページより大きい文書を提示する事で，ユーザが見ているページよりも，全般的に詳しい情報が載っていると思われるページを与える事が期待できる．

他にも，ユーザが気になる評価属性語が存在した場合は，その単語を指定してできるようにして重み付けユークリッド距離などを用いたり，コサイン相関値を用いる手法なども考えられる．

6. プロトタイプシステムの実装と考察

本提案手法の有効性を検証するため，実際に実験を行う．本実

順位	単語	登場回数	順位	単語	登場回数
1	ない	155	16	軽い	37
2	いい	120	17	厚い	33
3	薄い	117	18	重い	31
4	小さい	111	19	安い	29
5	良い	98	20	楽しい	23
6	大きい	90	21	にくい	23
7	やすい	86	22	すごい	23
8	欲しい	80	23	嬉しい	19
9	高い	74	24	長い	17
10	新しい	65	25	少ない	16
11	よい	58	26	近い	16
12	無い	52	27	早い	16
13	多い	45	28	美しい	15
14	悪い	44	29	ほしい	15
15	詳しい	41	30	っぽい	15
			31	⋮	⋮

表 1　登場回数の高い形容詞

験では，商品名はデジタルオーディオプレイヤー,「iPod nano」を用いた．

6.1 文書収集

Google [3] を用いて,「iPod nano AND (評価 OR レビュー)」というクエリで日本語のページを検索し，得られた検索結果から上位 100 件のページを取得し，タグを除去してテキストファイルに変換する．各文書ファイルの日本語解析については，形態素解析システム茶筌 [4] を用いて形態素解析を行った．

6.2 形容詞の頻度計算

100 個の文書集合から，出現する形容詞の頻度を求め，その出現頻度の多い順に並べる．表 1 にその結果の一部を挙げる．

取得した文書集合中で登場頻度が高い形容詞は，文書中の評判文において，評価属性語を形容する評価属性値に相当する形容詞である事が期待される．実験の結果,「薄い」「小さい」「軽い」と言った iPod nano の本体について，その「重さ」や「サイズ」などといった評価属性語を形容していると思われる形容詞や,「価格」を形容していると思われる「安い」といった単語が上位にあらわれている．

また「ない」「いい」「にくい」「やすい」と言った単語は，非自立な形容詞で，それのみでは何を形容しているかは分からないが，文中では

「(評価属性語) が〜〜しやすい（しにくい)」

などのように，他の語と結びついて，評価属性語を形容する評価属性値となっている事が多いと期待できる．

登場頻度が少ない形容詞については，評価属性値でない形容詞の可能性が高く，ノイズであると思われる．登場頻度のおよそ 30 位前後を境に,「弱い」「強い」「厳しい」と言った，ノイズと思われる形容詞が多く見られるように思われたため，本プロトタイプシステムによる実験では，登場回数の上位 30 件までの形容詞を用いて，評価属性語の抽出に用いる事にした．

6.3 評価属性語の抽出

登場頻度の高かった上位 30 件の形容詞を用いて，評価属性語となる名詞句を抽出する．100 個の文書集合に対して，登場頻度の上位 30 件の形容詞の位置を調べ，その近傍にある名詞を，形容詞に近い順から前後ともに c 個ずつ，取得していく．

「iPod nano は薄い」のように,「本体の大きさ（サイズ)」を意味する評価属性語が直接書かれていない評価文に関しても，次の文章，あるいは前の文章など，その近傍に，該当する評価属性語が書かれている可能性があるため，句点による文の区切りなどは考えず，形容詞の近傍の語句を取り出す．形容詞に近い順から前後 c 個ずつの語句の中にある名詞を調べて，その登場頻度を調べた．

c の値を 1 から漸増させて調べたところ，c=2，3 の時は精度は少しずつ上がっていったが，その後 c=4 以降は減少に転じた．これは，c の値があまり大きくなると，関係の無い文にまで名詞の抽出が及んでしまい，抽出の精度が落ちているのだと思われる．精度については，その計り方が難しい．

例えば，今回の iPod nano の例では「アップル」という名詞が上位に来る．この「アップル」という単語は iPod nano という製品そのものの性能に関連するような名詞では無く，製造・販売元の会社名である．商品の評価・評判情報を知りたいというユーザの中には，その商品を作っているメーカーも，自分の中での商品の評価を決める重要な一要素の人も居ると思われる．そういう人にとっては，例えば「アップルが製造している」という情報は iPod nano の評価・評判の対象とも言える情報であり,「アップル」は評価属性語だと考える事もできる．だが同時に，メーカーなど関係なく，その性能のみで評価を下す人も居る．

このように，抽出された各名詞句がその商品の評価属性語であるかどうかは，厳密には各人の主観による価値観も関わってくるため定義が難しい.「iPod nano の〜〜」と書いて意味が通じるものを評価属性語だと考える（この場合,「アップル」などの単語は除外され，性能などに関係する語句が中心になる）と，上で書いたように c=3 の時が一番精度が高く，上位 30 件について 17 個が評価属性語であると考える事ができる．

他にも,「アップル」のようなメーカー名や,「写真」(iPod nano では写真も見ることができるので，それについて言及した記事が多いと思われる) といった，iPod nano の評価につながると思われるような単語も評価属性語として考えると，21 個が評価属性語と考えることができた．すなわち，適合率は最大で 70%前後であると言える．

また，あまり登場頻度が下位になると，評価属性語では無い名詞句が多くなってくる．同時に，下位になるほど,「iPod nano の〜〜」という文章で意味が通じても,「使い勝手」や「こだわり」のような漠然とした名詞も増えてくる．そうしたノイズは上位 30 件前後を中心に目立つように思われたため，30 個を区切りとして考えた．

上位 30 件以降は評価属性語と思われる名詞句は少なくなっていくが，上位 31 件〜100 件の中にも，評価属性語とみなせる単語が存在する．100 件前後からは単語の登場頻度自体も 2，3 回となり，それ以降はほとんど評価属性語と見られる名詞は存在しなかった．上位 30 件までに入っている名詞句と同様の意味を表すとみなせるものは除き，上位 30 件の区切りによって

順位	名詞	登場回数	順位	名詞	登場回数
1	音	28	16	音楽	14
2	傷	28	17	手	13
3	液晶	25	18	ケーブル	13
4	ケース	21	19	パソコン	13
5	写真	21	20	順	12
6	価格	20	21	バッテリー	12
7	画面	20	22	本体	11
8	鉛筆	19	23	ボタン	11
9	サイズ	18	24	ビデオ	11
10	アップル	18	25	クリック	10
11	製品	17	26	ジャケット	10
12	幅	16	27	ディスプレイ	9
13	曲	14	28	デジタル	9
14	音質	14	29	重量	8
15	容量	14	30	文字	8
			31	⋮	⋮

表 2　評価属性語の抽出

名詞	共起度
液晶	1
カラー	0.3571429
搭載	0.1838235
情報	0.09210526
画面	0.08960573
ディスプレイ	0.08298755
デジタル	0.07920792
価格	0.07831325
プレーヤー	0.07692308
対応	0.07692308
モデル	0.07641196
発売	0.07520892
写真	0.0751634
機能	0.07219251
ビデオ	0.07210031
オーディオ	0.06976745
表示	0.06976745

表 3　評価属性語「液晶」の，全文書における文章中の共起度

切り捨てられていると考えられる評価属性語は「カラー」「質感」などおよそ 10 個前後であり，c=3 の場合再現率はおよそ 75%であると言える．

また，今回の実験では,「印象」「人気」「感じ」と言った，感想を表すような一般的な名詞は stop word として除外した．本プロトタイプシステムでは c=3 の，上位 30 件の名詞を iPod nano の評価属性語であるとして，次のステップに進む．この方法により取得された名詞とその登場回数を，表 2 に示す．

また，今回参考として，商品以外の対象についてはどの程度の精度があるのかを調べるため,「iPod nano」以外にも,「宇宙戦争」「京都大学」の例を用いて同様の実験を行った．結果としては,「宇宙戦争」の場合適合率が 60%，再現率が 65%で,「京都大学」の場合適合率が 55%，再現率が 35%と，精度は順に下がっていった.「iPod nano」のような電化製品などの商品は，ある程度評価属性語が明確に定まっており，どのユーザも同じようなポイントについて語る傾向があるため，精度は高くなっていると思われる．しかし,「宇宙戦争」のような映画作品の場合，レビューをする人は劇中の描写などについて語ることが多く,「地中」,「愛」,「家族」といった，それのみでは評価属性語とはみなすことの難しい名詞も頻度が高くなっていた．それでも，監督や俳優の名前,「迫力」「最後」と言った評価属性語とみなせる名詞も抽出できていた.「京都大学」のような組織名の場合,「学生」「教授」「環境」など，ある程度上位 30 件以内に評価属性語は抽出できているものの，評価属性語が多岐にわたっているため，再現率は低くなってしまった．また，一番の問題として，今回のページ収集の手法では「京都大学」の何かのポイントについてレビューしているような Web ページがあまり集められず，またそのようなページ自体数が少ないと思われるので，ノイズが多いことが挙げられる．

本手法がどのようなジャンルの対象に関して有効であるのかを詳しく調べるとともに，様々なジャンルの対象に関して有効になるような手法の改善について，今後考えていく必要がある．

6.4　文書の評価属性ベクトル計算

「iPod nano」の例に戻る．100 個の文書集合に対して，上で求めた 30 個の評価属性語について，各文書でどれだけ言及されているかをパラメータ化する．1 つの文書につき，各評価属性語の言及度のパラメータ 30 個からなるベクトルで表すことができるようにする．100 個の文書全体を解析し，その中の名詞句，形容詞句，形容動詞句，動詞句 (非自立なものは除く) 全てについて，30 個の評価属性語と同じ文章で共起している度合いを，Jaccard 係数を用いて求める．一例として，評価属性語「液晶」に対する共起度の一部を，表 3 に示す.「液晶」という単語は評価属性語そのものなので，共起度は 1 となっている．形容詞なども単語の対象であるが，上位は名詞で占められる結果となった．

このようにして，30 個の評価属性語について他の単語との共起度が求まる．次に，1 つ 1 つの文書の中の，それぞれの単語の出現数を求めて，その単語の共起度 × 単語の出現数だけ，対応するパラメータに数値を加えていく．

例えば，ある文書のベクトルを $(V_1, V_2, \cdots, V_p)$ として，その文書中に「カラー」という単語が 10 個あったのなら,「カラーは」「液晶」との共起度が 0.3571429 であるので，液晶に対応するパラメータ V_1 に 0.3571429×10 を加える.「カラー」という単語が他の評価属性語とも共起しているのなら．その対応するパラメータに同じようにして数値を加える．これを対象となる全ての単語について繰り返し行う事で，文書をベクトル化する事ができる．

6.5　ユーザに内容の異なるページを提示する

文書がベクトル化できたので，Google の上位 100 件の検索結果から，ユーザが 1 つページを選択した際に，そのページと最も内容の異なるページを提示する事を考える．本来はユーザは Google の検索結果から Web ページを選択する事を想定しているが，本プロトタイプシステムでは，文書集合の中から直接 1 つの文書を選択する事にした．

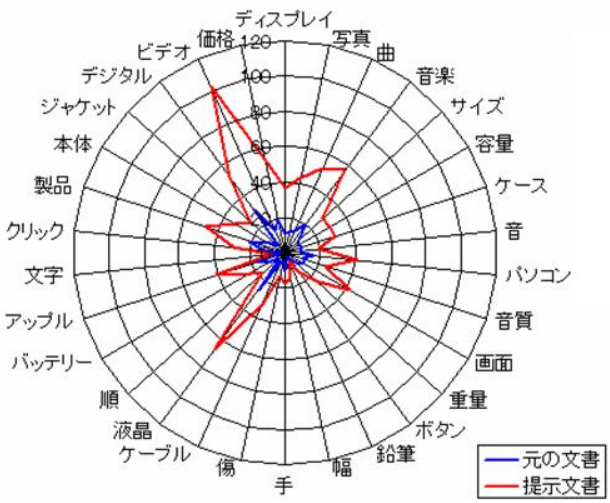

図 2　二つの文書の各評価属性語に対する言及度

文書間の距離の測り方については色々考えられるが，今回は単純にユークリッド距離を用いて，最も遠い文書を提示するようにした．今回はプロトタイプなので，検証のため，文書を提示するとともに，クエリである元のページと，システムが提示したページのベクトルも出力させ，数値での言及度の差が，実際の文書の内容をどれだけ反映しているかを考察できるようにした．

6.6　結果と考察

6.6.1　結　　果

1 つ文書を選んだところ，システムはその文書をクエリとして，その文書と比べて，評価属性語全般について言及度が高いと思われる他の文書を提示した．図 2 に，入力文書と出力文書各々における，評価属性語の言及度をレーダーグラフで比較したものを記す．今回の場合，ユークリッド距離を用いているため，結果としては，入力文書より全ての評価属性語において詳しく言及されている文書が答えとして提示された．

6.6.2　考　　察

これらのベクトル化された数値が，実際の文書における，各評価属性語の言及度を正確に反映しているのかを検証してみる．

例えば「サイズ」という評価属性語について，実際に二つの文書を見比べてみる．元の文章では「小さい」という事しか書かれていないが，提示された文章では正確な寸法などについても言及されている．また,「液晶」という評価属性語については，元の文章では記述が無かったが，提示された文章では，その大きさや視認性についての記述があった．

抽出した 30 個の評価属性語の中には,「手」や「鉛筆」，そして「製品」など，iPod nano の評価属性語では無いと思われるノイズも混じっていて，それらについてはどちらの文書がより多く言及されているかなどは判断する事ができない．その上で，30 個中 18 個，60%の評価属性語について，より多くの記述があると判断できた．この結果から，本提案手法で数値化した「言及度」は，実際の各文書における各評価属性語の言及の度合いに対して，ある程度の指標になりうると考えることができる．

参考として，出力として提示された最も距離が遠い文章以外の，距離の遠さにおいて上位にきていた他のページを調べたところ，提示された文書のように，クエリとなる文書よりも，より多くの評価属性語についてより詳しく書かれているようなページが多かったが，中には評価・評判情報が載っている Weblog などの，リンク集のようなページも含まれていた．このようなページは，その Web ページ自体に評価・評判情報があるわけではないが,「iPod nano」という言葉が含まれる Weblog などをリンク集のような形で集め，その本文の一部も snippet のような形で表示しているので，言及度が上がってしまったと考えられる．

また，今回実験に用いた「iPod nano」については，発売元のアップルが，同系統のシリーズとして「iPod」,「iPod Shuffle」なども発売しており，それらのレビューページも検索結果に混ざっている．これは今回用いた「iPod nano」の例に特有の問題とも考えられるが，このようなノイズを除く事も考えなければいけない．

7.　今後の課題とまとめ

7.1　今後の課題

7.1.1　ページ例の集合について

ページの収集に関しては，本研究の主題である「似て非なる検索」そのものの範囲外であるので，今回は非常に簡単な手法によりページ例を集めたが，人手によって先により正確なページ集合を作るなどの手法をとった方が,「似て非なる検索」の効果はよりわかりやすくなったと思われる．ページ収集のより良い手法については，今後の課題としたい．またページ例の本文から，例えば Weblog で言えばサイドバーやトラックバックの部分などのノイズを除き，対象についての評価を述べている部分のみを抽出する事も考えなくてはいけない．

7.1.2　評価属性語の抽出手法について

日本語の係り受け的な構造に着目した手法で抽出を行うなどすると，より商品の評価属性語の抽出の精度が上がる事が期待される．今後の研究ではそれらの手法も用いて評価属性語抽出の精度の向上を目指し，より「似て非なる検索」の効果がわかりやすくするため，その前段階の評価属性語の抽出に対しても，改良を検討していく必要がある．

7.1.3　評価属性語について

今回のプロトタイプシステムによって，iPod nano に関して抽出された評価属性語は，人間が見ると同じ物を指している物もある．例えば「液晶」「画面」などと言った単語は，今回例として使った iPod nano という商品に関しては,「画面が傷つきやすい」「液晶が傷つきやすい」と言ったように，実際の Web ページでは同じ意味で使われている事が多い．

また,「傷」という名詞は，iPod nano という商品に関しては評価属性語として抽出されたが，上の例でも分かるように,「画面」や「液晶」，あるいは「本体」などについての評価属性値の一部として出てくる事が多かった．

このように．本研究の評価属性語抽出手法で抽出された評価属性語どうしは，必ずしも並列的に独立して扱えるものとは限

らず,「液晶」と「画面」のように同列であったり,「液晶」と「傷」のように，一方の評価属性語が，もう一方の属性語に関連する語句であったりする場合が多く，独立して扱うのでは無く，それらを加味した評価属性語の処理ができれば，評価属性語抽出語の処理において，改良が望めると考えられる．

7.1.4 文書の提示手法について

今回の提案手法のような，ユークリッド距離を用いて，クエリとなるページと距離が最も遠く，かつ絶対値がそのページより大きい文書を提示する手法では，処理を繰り返して連続的な閲覧を行うことができない．すなわち，あるページをクエリとして「似て非なる」ページが出力として提示された時その提示されたページをクエリとしてさらに「似て非なる」検索を行うことができず，実質的には「最も多くの評価属性語について，最も詳しく書かれていると思われる文書」が優先的に提示されてしまう形となっている．

実際には，ユーザはページの連続的な閲覧も行うと思われるので，単純に全評価属性語についてのユークリッド距離を用いるのではなく，各文書が各評価属性語についてどの程度言及されているのかをより詳細に分析し，他の評価属性語について書かれていなくとも，ある評価属性語について詳しく載っているページなども提示対象となるような手法が必要である．

7.2 ま と め

本研究では, ある商品の評価・評判情報の存在するページ集合を用意し，その中の 1 つのページをクエリとして，ページ集合の中から「似て非なる」他のページを見つけ出す事を目的とした．その際，商品の評価対象となるトピックである評価属性語に注目し，その言及の度合いで文書をパラメータ化し，クエリのページに無いような言及がなされているページを探す事を考えた．

本研究では，ユーザが 1 つページを選んだ時，それに対して「似て非なる」ページを提示するという，ユーザの閲覧動作に連係してさらに情報を補完するようなページを閲覧できるようにするような手法を提案した．しかし，ある商品について，ユーザが多様な観点からの評価・評判情報を得るという目的を考えると，他の手法も考える事ができる．たとえば，ユーザが商品名のみを入力することで，検索エンジンを用いて適当なクエリで検索を行い，結果の上位 100 件などをページ例として取得する．そして．それらのページ集合を用いて，各ページについてどの評価属性語に関して言及されているかでグループ分けするようなクラスタリングを行い，各クラスタの代表ページを見るといった方法も考えることができる．この方法では，クラスタリングを具体的にどのように行うかが課題となると思われる．本論文においては，時間の都合上そのような他の手法についての検証を行うことができなかったが，今後は，こうした他の手法についても検証し，本提案手法との違いなどを考えていきたい．

また，本研究のプロトタイプシステムでは，GUI を作らず，コンソールによるプロトタイプシステムを用いて実験を行ったが，ユーザの知識発見という目的のため，ユーザインターフェースなども含めたシステム構築法も検討していきたい．

最後に，商品の評価・評判情報については，他にも「同じような評価属性語について言及されている」という同一の特徴を持っている文書の中で，商品名が異なるという「似て非なる検索」も考えられ，こちらの場合，ユーザがある商品の評価・評判情報のページを見ている時に，ページの内容から他のメーカーの同一ジャンルの商品をシステムが自動的に探し出して推薦するシステムへの応用が考えられる．今後は，上記で述べた課題点について考えるとともに，本研究で抽出された評価属性語を基に，同じ評価属性語を持つ別の商品を探す手法も考えていきたい．

謝　　辞

本研究の一部は，21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」，文部科学省科学技術振興費プロジェクト「異メディア・アーカイブの横断的検索・統合ソフトウェア開発」(代表:田中克己)，および，平成 17 年度科研費特定領域研究 (2)「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」(課題番号:16016247，代表:田中克己)，および，平成 17 年度科研費若手研究 (B)「参照の同一性判定に基づく複数 Web ページの検索閲覧方式の研究」(課題番号:16700097，代表:小山聡) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] Amazon.co.jp
http://www.amazon.co.jp/.

[2] 価格.com
http://kakaku.com/.

[3] Google
http://google.co.jp/.

[4] 形態素解析システム茶筌
http://chasen.naist-nara.ac.jp/.

[5] 立石, 石黒, 福島：“インターネットからの評判情報検索”, 情報処理学会研究報告, **2001-NL-144-11**, pp. 75–82 (2001).

[6] 鈴木, 高村, 奥村：“Semi-supervised な学習手法による評価表現分類”, 言語処理学会第 11 回年次大会 (2005).

[7] 藤村, 豊田, 喜連川：“文の構造を考慮した評判抽出手法”, 電子情報通信学会第 16 回データ工学ワークショップ (DEWS2005), **6C-i8**, .

[8] 熊本, 田中：“Web ニュース記事からの喜怒哀楽抽出”, 第 165 回 自然言語処理研究会.