データベースシステム 128－64
（２００２．７．１９）

検索語間の関連を考慮したWeb検索法の提案

大塚 真吾† 喜連川 優†

†東京大学 生産技術研究所
〒153-8505 東京都目黒区駒場 4-6-1

あらまし

Webページの利点の1つは気軽に情報発信が行える点があげられる。また、匿名性も高いため、各々のページの所有者は自分の地位や立場などに囚われずに、自分の思いを気軽に書くことができる。各々が自由に書いている意見を集め、それを解析することが可能になればWeb上での「世論」や「トレンド」を掴むことが可能となる。しかし、そのためには単語間の関連についても解析する必要があるが、現状の検索エンジンでは基本的に単語単位の検索しか行えない。そこで、本稿では日本語を含む大量のWebページに対してテキスト解を行い、様々な「意見」を抽出し、Webページ全体でのトレンドを見つけるための手法を提案する。

The Web Search Method for Efficient Relationship Between each Word

Shingo Otsuka† Masaru Kitsuregawa†

†Institute of Industrial Science, The University of Tokyo
4-6-1 Komaba Meguro-ku, Tokyo 153-8505, Japan

Abstract

The Web page has the advantage of dispatching Information lightheartedly and anonymity. Therefore, everyone can write opinions lightheartedly in the Web pages without worrying about his or her position and situation. If we collect and analyze the opinions which have been written by each person, we get public opinions and trends in the Web. Then, we need to analyze the relationship between each word. But the present search engines only search word-base basically. In this paper we can discuss the analyze of texts for a large number of the Web pages included Japanese and extract various opinions. And we also propose a method to find out trends.

1 はじめに

現在、Web上には大量のページがあり、今後も急激に増加すると予想される。Webページでは情報発信が気軽に行え、さらに匿名性も高いため各々のページの所有者は自分の地位や立場などに囚われずに、自分の思いを気軽に書くことができる。

最近では個人が訪れたさまざまなページについての批評をまとめた「Webログ」と呼ばれるものが人気を集めており多くのアクセスを集めている。また、日記や掲示板なども自分の意見を表現する場所になっており根強い人気を持っている。

これら個人の主観で書かれているページ情報や掲示板のログなどを大量に収集し、統計を取ることができれば、Web上での人々の意見やニーズ、すなわちトレンドをつかむことが可能になる。

しかし、一般のサーチエンジンを用いてキーワード検索を行う方法では効率が悪い。

そこで、本稿では日本語で書かれたWebページ全体からのトレンド抽出の提案を行う。また、本稿

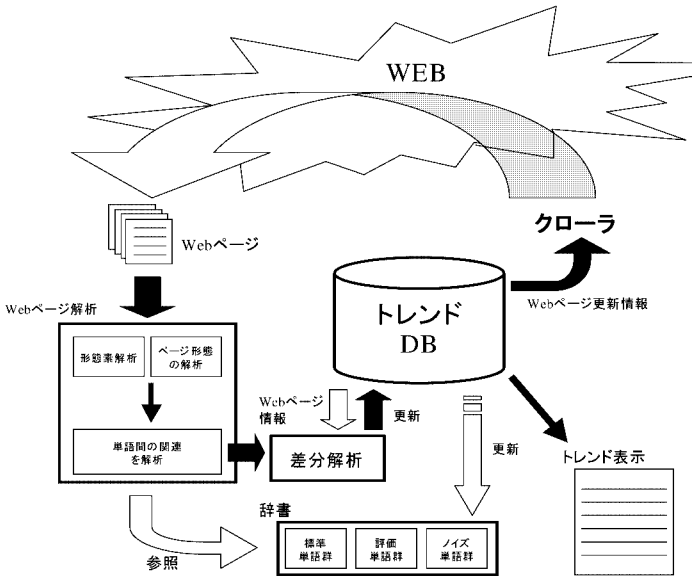

図 1: トレンド抽出の概要

では Web ページの変化を見るために、単語間の関連を利用した Web ページ解析法についも議論を行う。

2 関連研究

Web マイニングや Web クラスタリングなど、Web ページごとの特徴に着目した研究は数多くある [2, 5, 3, 8]。これらの研究では Web ページの類似性からページのジャンル分けを行っている。また、リンク情報からジャンル同士の関連なども分かる。しかし、Web ページを処理単位としているため、例えば掲示板や新聞記事の羅列など、ページの中の文章に主題がいくつもあると精度が下がる。また、ジャンルについてもあまり細かくすると、数が多くなり扱いにくい。これらの研究は、ユーザが興味を持つトピックの発見を支援することが目的であり、本研究と目的が異なる。

本研究ではある事柄の評価について着目するため、Web ページの中の文章に重点を置いている。同様な研究に単語間の関連を考慮した検索システムについての提案がある [4, 7]。しかし、単語間の関連を考慮した解析処理は、処理と保存のコストが非常に多く、Web 全体のような大量のデータに対して単語間の関連を考慮したページ解析手法は確立されていない。

また、Web ページからある特定の商品に関しての評価を探す研究も行われている [6]。これは、「良い、悪い」など評価に関する単語群をあらかじめ辞書に登録しておき、特定の単語と評価の単語群との距離から、その単語の評判を得るものである。この研究は、あらかじめ決められた商品を対象としているため、本研究の目的である Web 全体に対しての評価検索は行えない。

3 トレンド検索システム

本研究の目的は日本語などで書かれた Web ページ全体のトレンドを掴むことである。「トレンドを掴む」とは、「商品、人物、場所」などのある特定の事柄について「好き、おいしい、良い」などの評価や好みなどの傾向を得ることである。ここでは、評価に関する単語を「評価単語」、評価される単語を「被評価単語」と定義する。

人の嗜好は時間と共に変化するものであり、Web ページの所有者はその時の気分でページの新規追加、更新、削除を行う。例えば、今までイメージが良かった企業が、一晩で一変してしまうことなどは日常茶飯事である。トレンドの発見には過去のページと現在のページとの比較が重要である。以上のことを考慮したトレンド抽出処理の概念を図 1 に示す。処理は大きく分けて、

- Web ページの解析
- 過去のページとの比較
- トレンド DB
- 追加、更新、削除を考慮した Web ページのク

$$URL_1, t_n\{(被評価単語_1, 評価単語_1, \ldots, 評価単語_y), \ldots, (被評価単語_x, \ldots)\}$$
$$URL_2, t_n\{(被評価単語_1, 評価単語_1, \ldots, 評価単語_y), \ldots, (被評価単語_x, \ldots)\}$$
$$\vdots$$
$$URL_i, t_n\{(被評価単語_1, 評価単語_1, \ldots, 評価単語_y), \ldots, (被評価単語_x, \ldots)\}$$
(i = URL 数, n = 更新回数, x = 被評価単語の数, y = ある被評価単語に対する評価単語の数)

例：http：//www.xxx.com/index.html, t_1\{(鬼怒川, 楽しい, 安い), (日光, きれい, 涼しい)\}, t_2\{(鬼怒川, 楽しい, 高い), (日光, きれい, 涼しい)\}

図 2: データ構成

ロール

- 辞書
- トレンドの表示

に分かれる。クローラから集められた Web ページに対して、トレンドを抽出を行い、過去のページがある場合は比較を行い、結果をトレンド DB に格納するという手順である。

3.1 Web ページの解析

クローラが集めた Web ページから、被評価単語と評価単語を抜き出す。そのためには形態素解析、ページ形態解析、単語間の関連解析の 3 つの解析処理を行う。

形態素解析処理 日本語で書かれた文章は分かち書きがされていないため、形態素解析を行う。また、単語の活用形など、同じ単語でも語尾が変化するものについて、辞書を利用した解析を行い、活用形などの誤差をなくす。

ページ形態解析処理 Web ページの構成にはニュース記事、掲示板、日記など、いくつかの体系に分けることができる。ニュース記事などは特定の項目について述べているため、被評価単語と評価単語が比較的近くにあると予想できる。一方、あるテーマについての掲示板の場合、ページのトップに被評価単語がある場合が多く、評価単語との距離は遠い。この場合、タグの解析を用いて Web ページの構造を解析する必要がある。

単語間の関連解析処理 上記の処理結果を元に、被評価単語と評価単語の抽出を行う。また、ページ形態解析処理のタグ解析からリンク情報を得ることによって、被評価単語の重要度などの計算も行う。さらに、文脈の解析も行い、否定文ならば評価を反転させるなどの処理を行う。

3.2 過去のページとの比較

解析した Web ページの URL はトレンド DB 内に保存されている。解析された Web ページが DB 内に登録されている場合、そのページは更新されたことを意味する。

通常の検索エンジンなどは、古いページの情報を破棄してしまうケースが多いが、本研究では過去のページと現在のページの相違を見ることが重要なため、過去のページ情報も保持する。

トレンド DB に URL が登録されていない場合でも、更新ページと変わらない場合がある。例えば、会議の議事録や日記などは、更新を行わずに過去の記録として残しておき、新たなページを作成する場合である。この場合、URL は異なるが、基本的には同一ディレクトリにある。また、過去の記録を残しているページはページ構成が似ているケースが多い。したがって、データベースに登録されていないページでも、URL とページ構成が類似していれば更新と同様なケースとみなすことができる。

3.3 トレンド DB

データベース内には、Web ページの URL と評価単語、被評価単語が登録される。データの構成を図 2 に示す。被評価単語は 1 つ以上の評価単語との対になる。また、t は DB に登録した時刻のタイムスタンプで、最大値が最新の情報を表す。

例では、http://www.xxx.com/index.html の「鬼怒川」に対する評価が「安い」から「高い」に変化したことを表している。この例では、ある時点での評価について全て格納しているが、差分だけを格納することも可能である。

3.4 追加、更新、削除を考慮した Web ページのクロール

トレンドを掴むには、更新が多いページを重点的に集める Web ロボットの設計を行う必要がある。そこで、トレンド DB 内の更新情報から、収集を行うページの優先順位を決定する。また、Web ページの更新は、噂のように発信源から徐々に伝わっていくと仮定すると、Web ページの更新記録を解析することによって Web ページ更新の順番をある程度予測することできる可能性がある。新規ページの発見については既存の Web ロボットの利用や検索エンジンを利用する。

3.5 辞書

辞書は Web ページの解析処理で利用し、標準単語群、評価単語群、ノイズ単語群の３つからなる。標準単語群は形態素解析処理を行うときに利用する単語である。新たな単語が出現した場合は辞書に追加する。

評価単語群とノイズ単語群は単語間の関連解析処理で利用する。ノイズ単語とは、「は」「が」「である」など、多くのページに存在するものや、意味を成さないものを集めた単語群である。

また、流行語や死語があるように、単語の評価は時間によって変化する。そこで、トレンド DB の情報を利用して定期的に単語群の更新を行う。

3.6 トレンドの表示

トレンドの表示はトレンド DB 内の Web ページの更新記録と評価単語を用いて解析を行い、

1. ある時点からある時点までのトレンド表示
2. トレンドの移り変わりの表示

の 2 つがある。また、評価単語だけでなく、被評価単語についても傾向を見ることでトレンドが分かる場合もある。例えば、流行語などはある時点から突然増加するので、被評価単語の増加率についても解析を行う必要がある。

4 単語間の関連を考慮したページ解析

Web ページからトレンドを抽出するには、テキストの解析が重要である。テキスト解析法は大きく分けて、

1. 単語間の関連（係り受け）を解析する
2. 単語の頻度から重要度を計算する（tf*idf 法など）
3. 単語の共起や分布を解析する

などがあげられる。1 は処理の精度が高いが処理に時間がかかるため、大量な Web ページに対して行うのは不可能である。2 と 3 は Web クラスタリングなどで、大量な Web ページに対して実時間で処理を行うことができる。

本研究で行うテキスト解析は、評価単語と被評価単語を探せば良く、構文解析ほど複雑な処理を必要としない。

4.1 評価実験

単語間の関連の変化と被評価単語の増加率を掴むため、現在の Web ページと過去ページとの比較を行った。実験では、形態素解析に茶筅 [1] を用いた。茶筅の解析結果には名詞、助詞、形容詞など、品詞情報を付加することができる。今回は被評価単語を名詞とし、評価単語を形容詞と副詞にした。また、ページ内の評価単語から前方 20 単語と後方 10 単語以内にある被評価単語は関連があるとした。

実験に用いたデータは、

1. ある会社の社長の会見記事
2. 特定の社長について書かれたページ
3. 野球ファンの日記

である。1 は昨年と今年の同じ時期に行われた 5 回の会見記事の比較を行った。図 3 に結果の一部を示す。去年の記事は法案や規制などに関連したものが多い。また、それらの単語は「欲しい」との関連が強いことから、法律についての要望があることが分かる。一方、今年の記事は有線、伝送、高速などが多く、「新しい」との関連が強かった。

次に、特定の社長について書かれたページについての比較を図 4 に示す。これは、ある会社の社長について検索エンジンで検索を行い、上位の結果について就任前と現在とで比較を行った。就任前の方が年功序列や再建などあまりイメージが良くない単語が多いが、現在は増資や成功などかなりイメージが良くなっていることが分かる。

最後に、熱狂的な野球ファンが書いている日記について、昨年の 3〜5 月までと今年の同時期との比較の結果を図 5 に示す。昨年は選手個人についての批評が非常に多いのに対し、今年は連勝や勝利などチームの動向についての関連が高いことが分かる。この結果から、昨年と比べ、今年はチームの成績が良いことが分かる。また、成績が悪いと個人の批評

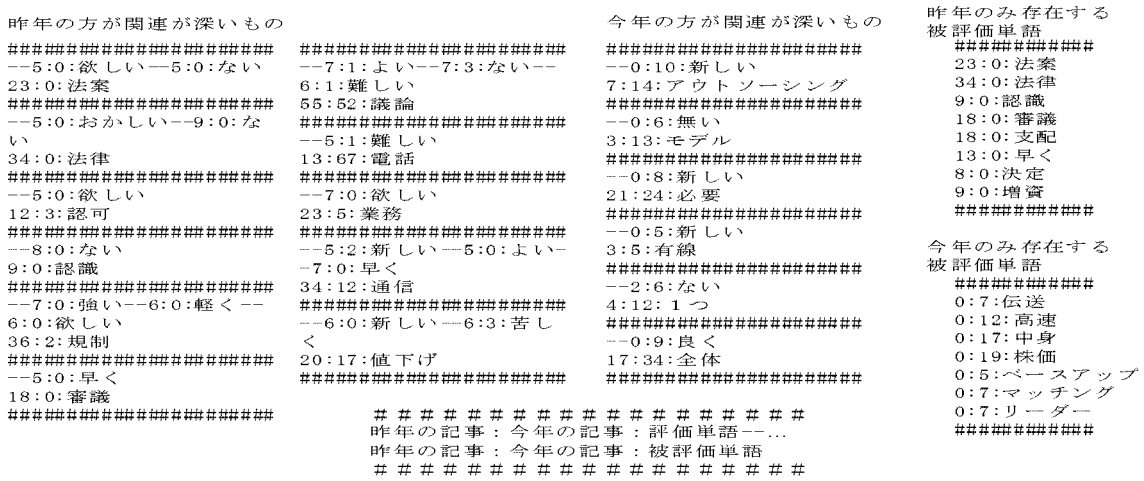

図 3: 社長の会見記事の比較

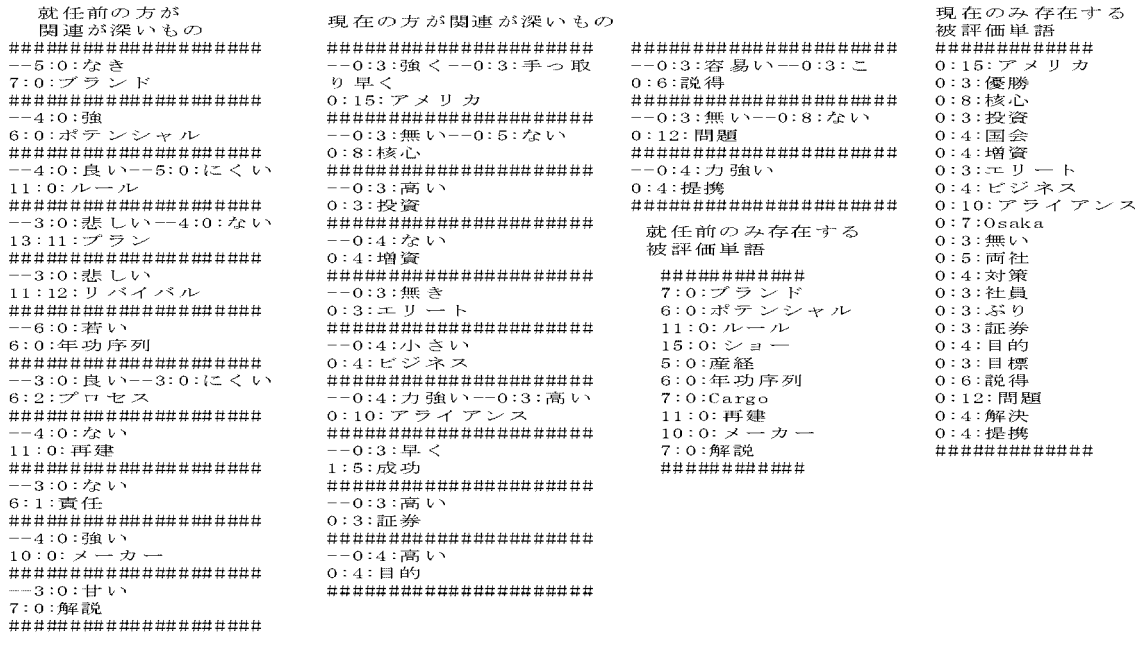

図 4: 特定の社長について書かれたページの比較

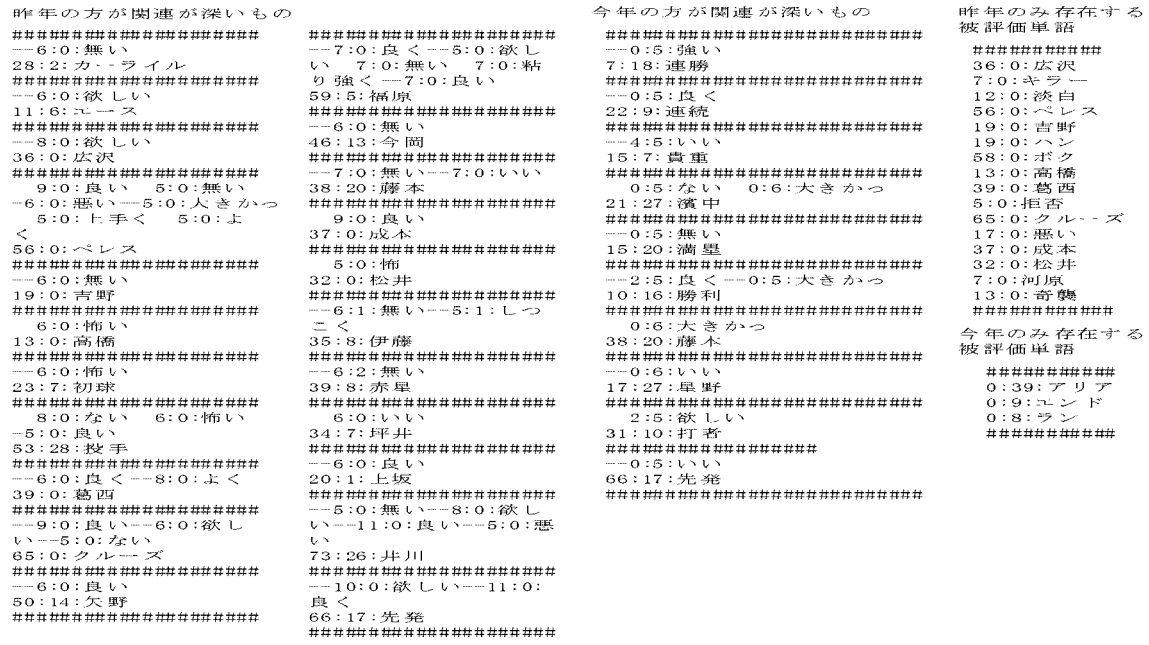

図 5: 野球ファンの日記の比較

を行う傾向があると予想できる。

このように、記事や日記など文書構造があまり複雑ではない Web ページに対しては、名詞、形容詞、副詞の関連を解析するだけでも、トレンドの移り変わりについて大まかに見ることができる。

5 おわりに

本稿では、Web 全体からトレンドを抽出するための概念について述べた。ある事柄に対する評価は時間と共に変化するものであり、その時々によって異なる。そこで、本稿では Web ページの更新に焦点を当てた解析方法について述べた。

さらに、本稿では Web ページの解析方法について実験を行い、ページ更新による評価の移り変わりを掴むことができた。また、今回は記事や日記など関連が分かりやすいページで実験を行ったが、掲示板などタグ構造の解析が必要なページについても実験を行う必要がある。

今後はこのシステムのプロトタイプを作成し、評価を行いたい。

参考文献

[1] 形態素解析ソフト「茶筅」http://chasen.aist-nara.ac.jp/.

[2] 福地健太郎, 豊田正史, 喜連川優 : Web Community Browser: 大規模 Web コミュニティチャートの可視化, 電子情報通信学会第 13 回データ工学ワークショップ (DEWS2002) (2002).

[3] 河野浩之, 川原稔 : Web 検索におけるテキストマイニング, 人工知能学会誌, Vol. 16, No. 2, pp. 212–218 (2001).

[4] 松村敦, 高須淳宏, 安達淳 : 単語の係受け関係を用いた情報検索手法の評価, 情報処理学会論文誌 : データベース, Vol. 41, No. SIG1(TOD5), pp. 22–30 (2000).

[5] 坂本比呂志, 有村博紀 : ウェブ・マイニング, 人工知能学会誌, Vol. 16, No. 2, pp. 233–238 (2001).

[6] 立石健二, 石黒義英, 福島俊一 : 評判情報検索システムの試作と評価, 情報処理学会第 63 回全国大会, Vol. 3, No. 27 (2001).

[7] 富田準二, 竹野浩, 菊井玄一郎, 林良彦, 池田哲夫 : グラフモデルの提案とテキスト検索システムへの適用による評価, 情報処理学会論文誌 : データベース, Vol. 43, No. SIG2(TOD13), pp. 94–107 (2002).

[8] 豊田正史 : WWW における関連コミュニティ群の発見, データベースシステム研究報告, Vol. 2000, No. 69, pp. 307–314 (2000).