Vol.2011-IS-118 No.1
2011/12/5

ユーザ嗜好に着目した評判情報の抽出手法に関する提案と評価

松澤　祐太[†]　　鈴木　裕利[†]
石井　成郎[††]　　小出　周之[†††]

インターネット上に存在する情報を収集，分析することは，企業にとってリスク管理，様々な傾向の調査等に利用が可能であり，個人にとっては商品購入時などの意思決定，選択の支援となり，非常に重要である．そこで，本研究では利用者が求める評判情報の抽出を行うため，利用者が編集可能な評価表現辞書を用いて，「ユーザ嗜好に着目した評判情報の抽出手法」の提案を目的とする．さらに，改善を図った抽出手法の有効性の向上について確認をするために実験を行い，その結果からは，処理時間の削減と抽出精度の向上が確認された．

Proposal and Evaluation of an Extraction Method of Reputation Information with Focusing on User's Preference

Yuta Matsuzawa[†] , Yuri Suzuki[†], Norio Ishii[††]
and Chikashi Koide[†††]

It is applicable to investigation of risk management and various tendencies etc. for a company to collect and analyze the information which exists on the Internet. It is also very important for an individual since it supports user's decision-making or selection at the time of merchandise purchase etc. So, it aims at making user's preference reflect in information gathering from the Internet in this research. Therefore, we propose the method of taking in the evaluation expression dictionary which a user can edit and specifically extracting reputation information. We experimented by implementing the proposal technique and checked the effect.

1. はじめに

近年，ネットワーク環境の普及に伴い，個人がパソコン，携帯電話などの端末を用いてインターネットに接続し，容易に情報を発信することが可能となっている．そして，利用者の増加に伴い，インターネット上には様々なコンテンツが増大し，多種多様な情報が存在する．個人が管理するコンテンツ，ウェブサイトも数多く存在し，そのようなコンテンツの代表としてブログが挙げられる．社会貢献や収益目的，あるいは自分の知見や情報を発信するためのブログには，様々な対象に関する評判情報や個人の主観的な意見が多く記載されているゆえに，これらの情報を収集，分析することは，企業リスクの管理，様々な傾向の調査等に利用でき非常に有益であるといえる．また，市場調査などのマーケティングやクレーム処理の支援においても活用が進んでいる．さらに，ブログには企業だけでなく，個人にとっても商品購入の事前調査等に役立つ情報が含まれているため，購入検討中の商品への他人の評価，意見が，意思決定，選択の支援となる．以上から，様々な分野，利用者にとってブログに記載されている個人が発信する評判情報の収集，分析は非常に重要であるといえる．

2. 関連研究

近年，Web 上に存在する文書に含まれる意見，評判を抽出する研究が盛んに行われており，ブログを扱った研究事例も増加している．対象とするテキストの単位によって，「評価情報を観点とした文書分類に関する研究」,「評価情報を含む文の抽出に関する研究」[3][4],「評価情報の単語単位での要素組の抽出に関する研究」[1][2]に分類される．

評判情報に関する研究として，立石らによる「インターネットからの評判情報検索」が挙げられる[1]．立石らの研究では，評判情報を「ユーザの行動，意思決定に役立つ形式で意見をまとめたもの」と定義して，特定の商品に対するユーザの行動，意思決定に有用な情報を提供する評判検索システムの構築を行っている．立石らはユーザが入力したキーワードとあらかじめ用意した評価表現辞書の評価表現を近接演算する方法を用いて，インターネット上に存在する Web ページから人の意見を抽出する手法を提案している．立石らのシステムは評判検索を行うために，Web ページ収集部，意見抽出部，分類・分析部の 3 つの機能で構成されている．Web ページ収集部は，ユーザが指定した商品に関する意見が記述されたインターネット上の Web ページをクローラを用いて収集する機能である．意見抽出部は，Web ページの文書の中からユーザが指

† 中部大学大学院工学研究科，春日井市
Graduate School of Engineering,Chubu University,Kasugai-shi,487-8501 Japan
†† 愛知きわみ看護短期大学，一宮市
Aichi Kiwami College of Nursing,Ichinomiya-shi,491-0063 Japan
††† ニフティ株式会社
NIFTY Corporation,Omori Bellport A, 6-26-1 Minami Oi, Shinagwa-ku, Tokyo 140-8544 JAPAN

定した商品に関する意見に該当する箇所を抽出する機能である．分類・分析部は意見抽出部で抽出した意見を「肯定・否定」に分類して検索結果として出力する．意見抽出部と分類分析部はユーザの行動，意思決定に有用な検索結果にするために重要な機能であると立石らは述べている．

立石らはこれらの機能を実現するために，評価表現辞書，パターンマッチングルールを用いている．さらに，意見抽出部で抽出した意見を「肯定・否定」に分類する処理を実現するために評価表現辞書に「肯定・否定」のラベルを付与している．

水口らの「Weblog を対象にしたリアルタイム評判情報分析システム eHyouban」[2]では不特定多数の対象の評判情報を検索可能にしている．eHyouban の大規模なブログ収集によるスケール感，ブログ記事から自動で評判情報を抽出するリアルタイム性を実現するために，ブログ大規模収集機能，ブログからの評判情報検索機能，ブログ／評判情報の検索結果を利用した記事数時系列変化や評判情報比較などの分析機能の 3 機能から構成されている．ブログ収集部では，ブログ更新情報(RSS 情報)を入力データとして用いており，常時ブログ記事本文を収集して，全ブログ記事の全文インデックスを作成している．同時に記事データを評判情報抽出部に渡している．評判情報抽出部では,記事データから評判情報を抽出し,評判情報インデックスを作成している．検索／分析部は，システム利用者の検索キーワードの入力によってキーワードを受け取り，ブログインデックスや評判情報インデックスを検索する．そして，ブログ検索結果と評判情報検索結果を HTML ページとして返している．

峠らの「ドメイン特徴語の自動取得による Web 掲示板からの意見文抽出」[3]では意見情報を抽出するには，人手による辞書の構築の負担の軽減や，辞書がドメインに依存してしまうことへの対処が必要と述べている．前述した立石らの抽出方法では評判情報を効率よく収集するため，利用者がドメインごとに辞書を人手により構築する必要があり手間がかかる．そこで峠らは大量の書き込みがある Web 掲示板から，抽出対象であるドメインごとに辞書を作成せず，ドメイン特徴語を自動取得し意見文を判別する手法を提案している．意見文の抽出方法として入力したテキストに現れた単語が意見文になりやすいか否かを学習し判定を行う．

橋本らの「階層非循環有向グラフを用いた文章の類似度に基づく評価文抽出」[4]では評価表現辞書を使用せずに，ドメイン依存／非依存に対応可能な評価文抽出手法を提案している．評価表現辞書を使用しない代わりに，人手により少量の評価文を与え，文章の類似度に基づいて評価文を抽出している． テキスト内の文法や意味的な情報を統一的に扱うことができる「階層非循環有向グラフ」と呼ばれるテキストの表現形式を用いることで高精度のテキスト処理タスクと，正確な類似度の算出が可能となっている．

以上，関連研究について言及したが評価情報の分析に関する各研究では評価情報の分析を評価表現辞書，パターンマッチングルール，形態素解析の技術を用いて評価情報の分析を行っている．以下では，各技術について説明する．

2.1 評価表現辞書

評価表現辞書は物事に対する人の評価を示す表現である評価表現の集合である．評価表現には「良い」，「最悪」などの人の感覚や感情を示す表現，「速い」，「うまい」などの物の性質や特徴を示す表現に加えて，「人気」「絶品」などといった名詞も該当する．各評価表現には「肯定・否定」のラベルが付与されており，各評価表現が肯定的であるか否定的であるか，極性が判断できるようになっている．この評価表現の例を表 1 に示す．

評価表現は物事に対する人の評価を示す表現であるため，対象となる物事の分野に大きく依存する．評価表現は分野固有の表現が存在し，評価表現は商品分野ごとに作成して辞書として用意しなければならない．さらに一般的に Web ページから人の作業により表現を収集して作成を行っているため，峠らは人手によるドメインに依存する評価表現辞書を作成せず，橋本らは評価表現辞書を使用しない手法を提案している．

表 1 評価表現の例
Table1　An example of evaluation expression

感情に関する評価表現		ものに関する評価表現	
評価表現	分類	評価表現	分類
良い	肯定	まろやか	肯定
使いやすい	肯定	おいしい	肯定
満足	肯定	いまいち	否定
高い	肯定	悪い	否定
最高	肯定	まずまず	肯定
悪い	否定	上品	肯定
お気に入り	肯定	強すぎる	否定
不満	否定	すばらしい	肯定
すばらしい	肯定	うまい	肯定
快適	肯定	好き	肯定

2.2 パターンマッチングルール

パターンマッチングルールは正規表現で記述され,適性値の計算に用いられる． この配列パターンから適性値を算出し，候補が相応しいか判断を行う．表 2 にパターンマッチングルールの例を示す．

表 2 パターンマッチングルールの例
Table2　An example of pattern matching rules

ID	パターンマッチングルール
1	商品名 .* (。|.|?|!) .* 評価表現	評価表現 .* (。|.|?|!) .* 商品名
2	商品名 . {0,12} 評価表現	評価表現 . {0,12}
3	商品名 .* (。|.|?|!) .* 評価表現	評価表現 .* (。|.|?|!) .* 商品名
4	商品名 .* 評価表現
5	評価表現 . {0,12} (、|。|.|!)
6	評価表現 .* か?

立石の研究ではパターンマッチングルールが 12 個用意され，抽出した意見に対し正規表現で記述した表 2 のようなルールを用いて適正値の判定を行っている．ルール数が 12 個であるため，12 次元の配列を用意し，抽出した意見が 12 個のルールそれぞれについてルールを満たす時は 1，満たさない時は 0 の値をセットする．この配列パターンから適性値を算出し，候補が適切かどうかについての判断を行っている．

2.3 形態素解析

文書や文より細かい語句単位で評価情報の抽出を行う研究では，形態素解析の技術を用いてマイニングを行い，評判情報の抽出精度を高めている．評判情報の抽出対象である Web 上の掲示板やブログなどのテキストは形式的に記述されておらず，テキストとしては質が悪い．さらに高い解析精度を得るため表記の多様性や局所的で特有な表現，略語などが多いため，これらに対応できる形態素解析の基礎言語解析技術が必要となる．

和多らの「単語の出現頻度に着目した病院評判情報の分析」[5]では，感性に関わる特徴を捉えるため，文書あるいは文書群の特徴を名詞，動詞，形容詞，副詞などの品詞ごとの単語の集合として捉える方式を提案している．Web 上に存在する評判情報に対し，形態素解析を行うことで品詞体系を求め，特定のテーマに関する評判情報の文書群と一般的文書群の単語の出現頻度を比較することにより，そのテーマの特徴的な単語を抽出している．和多らは Web 上の病院評判情報についての文書群と夏目漱石の小説の文書群を比較することで，病院評判情報についての特徴的な名詞，動詞，形容詞，副詞を求めている．

同様に鈴木らが行っている「Weblog を対象とした評価表現抽出」[6]においても形態素解析が使われている．評価対象・属性・評価語の三つ組みで構成されている評価表現を抽出して，文書全体が肯定的な評価であるか，否定的な評価であるかを判定している．評判情報を含む文の中でどの対象の，どの部分が，どうなのかという 3 要素から有効な評判情報を抽出し，評判情報を肯定的評価，否定的評価，非評価のいずれかに分類している．

3. 目的

評価情報の抽出はいろいろな方法があるが利用者の嗜好に合わせた情報を抽出する手法は少ない．本研究ではインターネット上に存在するブログから利用者が求める評判情報の抽出を行うため，利用者が編集可能な評価表現辞書を用いて，利用者の嗜好に合わせた評判情報の抽出をする．峠ら[3]は人手によるドメインに依存する評価表現辞書を作成せず，橋本ら[4]は評価表現辞書を使用しない手法である．しかし，本研究ではドメインごとに依存する辞書ではなく，利用者ごとに依存する辞書を利用者が編集や追加することで利用者にとって有用な評判情報を抽出する．さらに利用者による追加，編集が容易に行える評価表現辞書の構築を目指す．

4. 提案手法

本手法では，利用者の嗜好に合わせた評判情報の抽出をするため，評価表現辞書は利用者が編集を行うことが可能となっている．さらに，評価表現には重みが付与されており利用者が重みを設定することで利用者にとって重要な単語を見落とさないようにしている．評判情報を抽出する概要図を図 1 に示す．

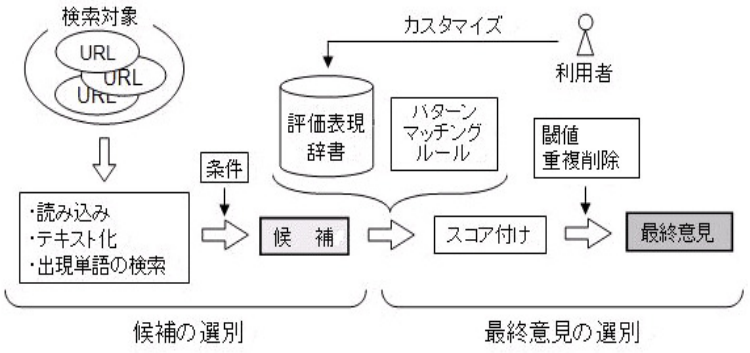

図 1 評判情報抽出の概要図
Figure1 Schematic diagram of information extraction reputation

候補の選別では，登録した検索対象である URL から文章を取得し，その文章を形態素解析用いて出現単語を検索し，各単語の出現頻度と単語の長さの条件を満たす単語を候補として採用する処理を行う．

最終意見の選別では，評価表現辞書とパターンマッチングルールを用いて候補の選

別で採用した各候補に対してスコア付けを行い，そのスコアが条件を満たす意見を最終意見を選別する処理を行う．

4.1 候補の選別

候補の選別では，検索対象である URL をあらかじめ登録する．登録された URL の Web ページのソースからテキストを取得して，形態素解析ツールを用いずにそのテキストに 2 語以上連続して出現した漢字を単語として抽出する．テキストから抽出した各単語の出現回数と単語の長さを求めて，候補の条件式からの条件を満たす単語を候補として採用する処理を行う．具体的な候補の条件を式(1)，(2)に示す．

URL から取得した文章に，利用者が入力したキーワードの単語が含まれている場合，利用者が入力した単語を含む文章の全てを最終意見として扱うと，利用者の行動，意思決定に有用な評判情報の検索結果として不必要な情報まで抽出してしまう可能性が非常に高いといえる．不要な情報の抽出を防ぐ方法として，候補を選別するための条件を設定し，その条件を満たした文章を候補として扱う．

$$(1.5 < TF_i \text{ and } TF_i < 7) \text{and } (1 < KEY_{LEN_i}) \quad (1)$$

TF_i ： 各単語の出現頻度

$$TF_i = \left(\frac{KEY_i}{\sum_k n_k}\right) * 100 \ [\%] \quad (2)$$

KEY_i ： 各単語の出現回数

$\sum_k n_k$ ： 文章に出現する総単語数

TF_iは文章における各単語の出現頻度を表し，式(2)のように求められる．また式(2)のKEY_iは各単語における出現回数を表し，$\sum_k n_k$は文章に出現する単語長 1 以上の総単語数を表している．

本研究では出現頻度TF_iの範囲を 1.5%から 7%とした．

4.2 最終意見の選別

最終意見の選別では，評価表現辞書とパターンマッチングルールを用いて候補の選別で採用した各候補に対して点数付けを行い，その点数が条件を満たす意見を最終意見として扱っている．以下に点数の算出式を式(3)に示す．

評価表現辞書は利用者の嗜好に合わせた情報を抽出するため，利用者が自由に編集を行うことが可能となっている．評価表現辞書には人の感覚や感情を示す表現，性質や特徴を示す表現が記載され，それぞれの評価表現には”肯定”・“否定”・“無属性”のラベルが付与されている．また，各評価表現には重みが付与され，最終選別において用いられている．利用者が重みを設定することで利用者にとって重要な単語を見落とさないようにしている．各評価表現に対する重みは 1~9 の間で設定でき，1 が最も重要性が低く，9 が最も重要性が高いことを意味している．評価表現辞書の例を表 3 に示す．

パターンマッチングルールは正規表現で記述され，最終意見の選別において用いられている．利用者が入力したキーワードとパターンマッチングルールの一致数で情報が適正かどうか判断している．パターンマッチングルールの例を表 4 に示す．

表 3 評価表現辞書の例
Table 3 An example of evaluation expression dictionary

評価表現	肯定/否定/無記入	重み
処分	否定	1
東京株式市場	−	1
いい加減	否定	4
弛む	否定	1
弛み	否定	1
立派	肯定	5
バカ	否定	5
ありえない	否定	4

表 4 パターンマッチングルール
Table 4 Pattern matching rules

1	_KEY_EXP_.*?。
2	_KEY_(の|が|は)(.*?_EXP_.*?)+.*?。
3	_KEY_(の|が|は).*?_EXP_.*?。
4	が.*?_EXP_.*?_KEY_.*?。
5	_KEY_.*?_EXP_.*?。
6	_EXP_KEY_.*?。
7	_KEY_.*?円.*?_EXP_.*?。
8	_KEY_(.*?円)+.*?。

KEY : キーワード　　EXP : 評価表現リスト

$$PAT\_COUNT * \left(\alpha TF_i + \sum_{i=1}^{n} \left(\frac{EXP\_WEI_i}{\beta}\right)\right) \quad (3)$$

PAT_COUNT : パターンマッチングルールに一致した個数

TF_i ： 各単語における出現頻度

EXP_WEI_i ： 文章中に出現する評価表現の重み

α，β ： 任意定数

PAT_COUNT はパターンマッチングルールの 8 個のパターンの中で，候補の単語が一致した個数を表しており，TF_iは候補の選別時に式(2)を用いて算出された，各単語における出現頻度を意味している．EXP_WEI_lは各候補の文章中に出現する単語の評価表現辞書に記載されている重みである．α，βは任意の定数を意味し，本研究ではα=1，β=100 と設定している．

各候補に対して点数付処理を行った後，点数に閾値を設定し最終意見の選別を行う．本研究では閾値を 5 以上 20 未満に設定し最終意見の選別を行っている．

4.3 評価実験

本章で示した提案手法を用いて抽出した評判情報が利用者にとって適正であるか評価するために実験を行い，その結果について考察をする[7]．70 件の株に関する評判情報を評価対象として用いる．この評判情報は(3)式から算出された点数が高い上位 70 件の評判情報となっている．株取引の経験を持たない被験者が 70 件の評判情報に対して，株取引において「必要な情報」であるか，「不要な情報」であるかの判断をしてもらう．この実験を 8 名の被験者に行ってもらい，70 件の評判情報のうち 4 名以上が「必要な情報」と判断した評判情報の件数と割合について調査する．それぞれ被験者は「必要な情報」，「不要な情報」と判断する基準が異なるが，「必要な情報」と判断された割合により利用者が必要だと思う情報が抽出できたかどうかを判断することができる．

この実験の結果，70 件中 34 件の評判情報が「必要な情報」と判断され 48.57%の抽出精度を得ているが，複数の問題点が挙げられている．第 1 に，抽出方法について改善の余地があり，特に候補に対しての点数付けの方法について検討する必要があるといえる．評価表現辞書を変更した場合に，検索に必要な情報の更新を行う必要があるが，更新にかかる処理時間が短い程，利用者は容易に評価表現辞書の変更を行うことができる．しかし，検索に必要な情報の更新に 126 時間の多大な時間を要しているため実用性に欠けている．さらに，データベースの検索に必要な際に必要な情報についてテーブルを作成して格納しているが，テーブル内には重複，類似している情報も存在しているために，検索結果も重複，類似したものが多くなっている．候補の選別で使用する形態素解析は PHP プログラムで行い出現単語の品詞体系が求められないので，2 文字以上続く「漢字」または「カタカナ」を単語と設定しているために，キーワードとして不適切な単語が候補として扱われている．このため 1 文字の単語の抽出も不可能となっている．

4.4 提案手法の改善

実験から得られた結果より提案手法の改善を行う．

候補の選別では，登録した検索対象である URL から文章を取得し，その文章を形態素解析用いて出現単語を検索し，各単語の出現頻度と単語の長さの条件を満たす単語を候補として採用する処理を行う．形態素解析を文章中に出現する単語の検索に用いるが，文章から検索する単語の抽出精度を向上させ，出現単語の品詞体系を求めキーワードとして最も適している「名詞」を候補として扱うために，形態素解析ツール Mecab を導入している．

形態素解析ツールの導入によって，1 文字から単語を検索できるようになったために，式(1)から単語の長さ 2 文字以上という条件である(1<KEY_LEN_l)を取り除いている．具体的な候補の条件として式(4) ，(5) に示す．

$$(1.5 < TF_i \text{ and } TF_i < 7) \tag{4}$$

TF_i ： 各単語の出現頻度

$$TF_i = \left(\frac{KEY_i}{\sum_k n_k}\right) * 100 \ [\%] \tag{5}$$

KEY_i ： 各単語の出現回数

$\sum_k n_k$ ： 文章に出現する総単語数

5. システム概要

本研究では第 4 章で述べた提案手法を前提に評判情報検索システムを構築する．本システムでは，検索に必要な情報を作成するため，キーワード検索を行う前に(1)~(4)の処理を行う．

(1)インターネット上に存在するブログの URL を収集し，URL を羅列した csv ファイルを準備する．URL を記述した csv ファイルから，データベース内の URL_LIST テーブルに URL を登録する．

(2)URL_LIST テーブルに登録した URL から Web ページを呼び出して，その Web ページのソースからテキスト情報のみを取得する．取得したテキスト情報をテキストファイルとしてローカルに保存する．

以上の(2)の処理を図 2 に示す．

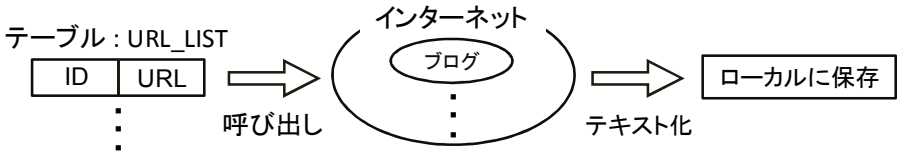

図 2　ブログに存在するテキスト情報の取得から保存までの処理
Figure 2　Processing from acquisition of text information that exists in blog to preservation

(3)ローカルに保存されたテキストファイルを読み込み，形態素解析ツールを用いて形態素解析処理を行い，URL の Web ページ内の文章であるテキストファイルから出現する名詞の単語の検索を行う．また，同時に Web ページのタイトルの抽出を行う．その後，単語の出現頻度を第 4 章で述べた式(5)から算出し，式(4)の条件を満たした単語のみを候補として扱い URL_TEXT テーブルに「URL」「タイトル」「単語」「出現頻度」の情報を登録する．以上の処理を図 3 に示す．

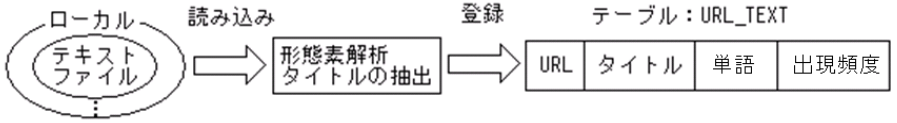

図 3　テキスト情報の読み込みから候補の選別までの処理
Figure 3　Processing from acquisition of text information that exists in blog to preservation

(4)(3)の処理から作られた候補のテーブル URL_TEXT を参照し，候補のテーブルの各単語とローカルに保存してある文章中の単語と合致する単語の前後 200 文字を切り出し，この文章を最終意見の文章として扱う．切り出された文章に対して，最終意見の選別処理を行い，各候補に対し第 4 章で述べた式(3)を用いて点数付けを行う．その後，設定した閾値を満たす点数を持った候補を URL_SCORE テーブルに「URL」「タイトル」「単語」「出現頻度」「文章」「点数」の情報を登録する．以上の処理を図 4 に示す．

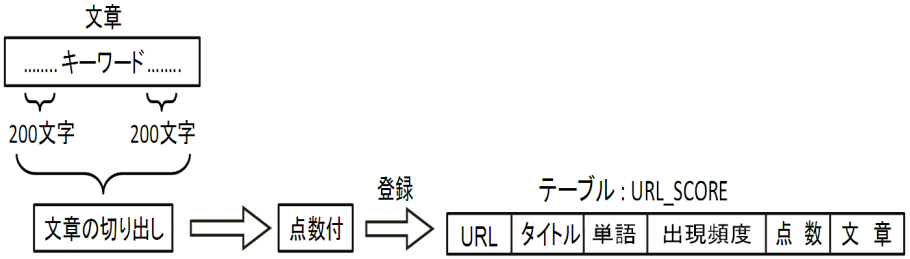

図 4　最終意見の文章の切り出しから採点情報の登録までの処理
Figure 4　Processing from cutting out sentences of the final opinion to registration of grading information

また，重複，類似した情報を除去するため，URL_SCORE に候補を登録する際に，「URL」「タイトル」「単語」の 3 つの要素が全て同じ情報がすでに URL_SCORE に登録してある場合は登録を行わない．

以上の(1)~(4)の処理をキーワード検索の前に行うことで，キーワード入力から検索結果の出力までの時間が短縮されて，評判情報の検索が可能となる．本システムでは利用者が入力したキーワードから PHP プログラムにより URL_SCORE テーブルにアクセスを行い，情報の取得を行っている．これらの情報の作成は PHP プログラムで行われ，全てブラウザ上で行われる．第 3 章で述べた候補の選別が(2)，(3)の処理にあたり，最終意見の選別が(4)の処理となる．実際に情報が登録された URL_SCORE テーブルを表 5 に示す．

表 5　URL_SCORE テーブル
Table5　URL_SCORE table

url	title	tags	score
http://102070.blog51.fc2.com/	保険の見直し	記事	17.1013
http://102070.blog51.fc2.com/	保険の見直し	視力	17.5013
http://102070.blog51.fc2.com/	保険の見直し	見直し	17.3413
http://102070.blog51.fc2.com/	保険の見直し	中	17.1813
http://102070.blog51.fc2.com/	保険の見直し	バック	17.1813
http://102070.blog51.fc2.com/	保険の見直し	視界	13.777
http://102070.blog51.fc2.com/	保険の見直し	期待	13.777
http://102070.blog51.fc2.com/	保険の見直し	昔	13.777
http://102070.blog51.fc2.com/	保険の見直し	日本	14.337
http://102070.blog51.fc2.com/	保険の見直し	内容	13.857
http://12blog27.blog74.fc2.com/	デイトレ ひとり旅	方	19.383
http://12blog27.blog74.fc2.com/	デイトレ ひとり旅	確	16.8312
http://12blog27.blog74.fc2.com/	デイトレ ひとり旅	無限	16.6712
http://12blog27.blog74.fc2.com/	デイトレ ひとり旅	東光	16.4312
http://12blog27.blog74.fc2.com/	デイトレ ひとり旅	割れ	17.3112
http://12blog27.blog74.fc2.com/	デイトレ ひとり旅	利	16.8312
http://12blog27.blog74.fc2.com/	デイトレ ひとり旅	下げ	16.8312
http://12blog27.blog74.fc2.com/	デイトレ ひとり旅	上	16.6712
http://12blog27.blog74.fc2.com/	デイトレ ひとり旅	ユーロ	16.8312
http://12blog27.blog74.fc2.com/	デイトレ ひとり旅	トヨタ	16.8312
http://12blog27.blog74.fc2.com/	デイトレ ひとり旅	岳	13.9594
http://12blog27.blog74.fc2.com/	デイトレ ひとり旅	トレード	14.3594
http://a-char.cocolog-nifty.com/	空売りで2億円を創っている日記	面	16.7266
http://a-char.cocolog-nifty.com/	空売りで2億円を創っている日記	金	16.7266
http://a-char.cocolog-nifty.com/	空売りで2億円を創っている日記	結果	16.8866
http://a-char.cocolog-nifty.com/	空売りで2億円を創っている日記	税金	16.7266
http://a-char.cocolog-nifty.com/	空売りで2億円を創っている日記	私	16.7266
http://a-char.cocolog-nifty.com/	空売りで2億円を創っている日記	毎日	16.8866
http://a-char.cocolog-nifty.com/	空売りで2億円を創っている日記	楽しみ	16.8066
http://a-char.cocolog-nifty.com/	空売りで2億円を創っている日記	日	16.8066
http://a-char.cocolog-nifty.com/	空売りで2億円を創っている日記	忠実	16.8066
http://agekabu.blog118.fc2.com/	明日上がる株を毎日予想するブログ	上	19.2615
http://agekabu.blog118.fc2.com/	明日上がる株を毎日予想するブログ	素人	12.6277
http://agekabu.blog118.fc2.com/	明日上がる株を毎日予想するブログ	初心者	12.6277

6.　実験

6.1　事前処理の速度実験

本システムでは，評価表現辞書の変更後に登録されている最終意見の更新を行う必要があるが，利用者が手軽に評価表現辞書の変更を行うためには処理時間が短いこと

が重要である．改善前のシステムでは事前処理に多大な時間を必要としている．評判情報検索に必要な事前処理にかかる処理速度と改善前のシステムの処理速度と比較することにより,事前処理の処理速度について向上しているかどうかについて確認する．

本実験は株に関する評判情報について抽出を行う．事前に，株について記述があるブログ 337 件の URL を URL_LIST テーブルに登録する. この 337 件の URL に対して，本システムと改善前のシステムそれぞれで処理を行い，その処理に要する時間を計測する．改善前のシステムの結果を表 6，提案システムの結果を表 7 に示す．

表 6　改善前のシステムにおける工程別のレコード数と処理時間
Table 6　Number of records and processing time according to process of measuring it with system before

	レコード数	処理時間[s]
(1)(2)の処理	10504	12.2[m]
(3)の処理	69839	110.3[h]

表 7　提案システムにおける工程別のレコード数と処理時間
Table 7　Number of records and processing time according to process of measuring it with present proposal system

	レコード数	処理時間[s]
(1)(2)の処理	11760	15.0[m]
(3)の処理	3655	3.8[h]

提案システムでは約 3.7 時間を要し，改善前のシステムでは約 110 時間を要した．生成された最終意見は，本システムでは 3655 件，改善前のシステムでは 69839 件であった．ことから，評判情報の検索を行うための事前処理に要する時間が減少したといえる．

(1)(2)の処理において, 改善前のシステムと比較して提案システムの生成されたレコード数は増加している．これは(2)の処理においてテキストファイルから出現する名詞の単語の検索を行う際に，本システムでは形態素解析を行うために形態素解析ツールを用いているためだと考えられる．改善前のシステムでは 2 文字以上の単語の抽出しか行えなかったが，本システムでは形態素解析ツールを用いることにより 1 文字の単語から抽出することが可能になったために, (1)(2)の処理の候補として生成されたレコードの数が増加している．

事前処理に要する時間を減少させた主な要因として，(3)の処理で生成された最終意見の削減が挙げられる．改善前のシステムでは(3)の処理で重複，類似した情報の最終意見が多量に登録されるため，処理にも多大な時間を要していた．本システムでは第 5 章で述べたように情報を登録する際に重複，類似した情報がすでに URL_SCORE テーブルに登録されていた場合，その情報の登録を行わない．よって，類似，重複した情報の削減により(3)の処理でのレコード数が減少して, (3)の処理に要する時間も減少した．

6.2 検索結果の比較実験

本実験では改善したシステムで抽出した情報の検索結果と改善前のシステムで抽出した情報の検索結果を比較して，検索結果の精度が向上しているかについて確認を行う．各システムで生成した最終意見に対してキーワード”株式”とし，株に関する評判情報の検索を行う．

改善前のシステムでの検索結果は上位 100 件において重複率が 90%となっており，100 件のうち 90 件の検索結果が類似している．本システムでは，21 件の検索結果の URL，タイトル，文章が異なり重複は確認されなかった．以上より，改善前のシステムに比べ利用者にとって実用性のある検索結果となったことが確認された．

7. おわりに

本文では，本研究の提案する評判情報検索システムについて，その特徴と実験結果について報告した．実験は，提案システムの改善の効果を確認するために，評判情報の検索に必要な情報の作成を事前処理として比較した．そして，実験の結果，事前処理に要する時間が，本システムの改善によって約 100 時間の削減となったことが確認された．また，提案システムによって抽出される情報の内容についても改善の効果を比較した. 改善前のシステムで抽出を行った情報の検索結果は, 類似, 重複したタイトル，文章が出力されているが，改善システムで抽出した情報では，タイトル，文章と URL がすべて異なっており，利用者にとって実用性のある検索結果となったことが確認された．

本文で報告したように，本手法では評価表現辞書を利用者が編集することが可能だが，評価表現辞書を変更した場合に，評判情報の検索に必要な最終意見の更新を事前処理として行う必要がある．そのため，利用者が効率的に評価表現辞書の変更を行うためには処理時間が短いことが重要である．よってシステムの改善は有用であったといえる．しかし，登録する URL の件数増加に伴って，事前処理に要する時間の増加は容易に予測される．よって，利用者にとってより利用が容易なシステムを構築するためには，さらに，処理時間の削減を行う必要があるといえる．同時に，登録する URL の増加によって抽出される情報の内容がより利用者にとって適正であるかについての確認も必要である．今後は，これらの点について，システムの改善，実験による検証

Vol.2011-IS-118 No.1
2011/12/5

を進めていく予定である.

参考文献

1) 立石健二，石黒義英，福島俊一:インターネットからの評判情報検索，人工知能学会誌 19 巻 3 号 ，pp.75-82 ，2004
2) 水口弘紀，槌田正明，久寿居大:Weblog を対象にしたリアルタイム評判情報分析システム eHyouban，電子情報通信学会第 19 回データ工学ワークショップ・2008 第 6 回日本データベース学会年次大会予稿集，Vol.DEWS2008，No.12-27，2008
3) 峠泰, 大橋一輝, 山本和英:ドメイン特徴語の自動取得による Web 掲示板からの意見文抽出，言語処理学会第 11 回年次大会発表論文集，pp.672-675，2005
4) 橋本大吾，嶋田和孝，遠藤勉:階層非循環有向グラフを用いた文章の類似度に基づく評価文抽出，橋本大吾，言語処理学会第 14 回年次大会論文集，pp.725-728，2008
5) 和多太樹, 関隆宏, 田中省作, 廣川佐千男:単語の出現頻度に着目した病院評判情報の分析，情報処理学会研究報告，Vol.SLP-2005，No.50，pp.15-20，2005
6) 鈴木泰裕，高村大也，奥村学：Weblog を対象とした評価表現抽出，人工知能学会研究会資料，Vol.SIG-SW&ONT-A401，No.2，pp.1-10，2004
7) 評判情報の抽出方法に関する一考察:松澤祐太，鈴木裕利，石井成郎，小出周之，電気関係学会東海支部連合大会講演論文集，ROMBUNNO.G3-6，2010