企業評判情報の分析に有効な風評テキストマイニング技術

Advanced Text Mining Technology for Corporate Reputation Information

櫻井　茂明
■ SAKURAI Shigeaki

インターネット上には多数の掲示板サイトがあり，特定の話題について互いに意見が交換されている。この話題の中には企業活動に風評被害を与えるものも存在している。

東芝は，掲示板サイトにおける記事のまとまり（以下，スレッドと言う）の中から，特定の企業に大きな風評被害を与えるおそれのあるスレッドを，早期にかつ自動的に抽出する風評テキストマイニング技術を開発した。自然言語処理技術やテキストマイニング技術などを組み合わせ，当社独自の方式によりスレッドを特徴付け，注意すべきスレッドやその中で話題になっている問題に関する表現を抽出できる。実際の記事を利用した抽出結果は，人間系によるものと高い割合で一致した。

Toshiba has developed a technology that makes it possible to automatically discover, at an early stage, important threads that might cause significant damage to a particular corporation or organization from sets of articles related to specific topics on bulletin board sites. Using both text mining and natural language processing techniques, this technology performs original characterization of threads and can extract important threads and expressions related to topics in the threads using these characterizations.

We evaluated the effectiveness of the newly developed technology using articles collected from bulletin board sites, and confirmed that the results based on the technology corresponded to user-based results with high probability.

1　まえがき

インターネット環境の普及に伴って，掲示板サイトを介した記事の発信が容易になり，多数の人々が互いの意見を気軽に交換できるようになっている。特定の話題について話し合われているこのような記事のまとまり（以下，スレッドと言う）の多くは，人々の関心を引くようなものではないが，中には企業活動にまで風評被害を与えるものも存在している。このような風評を放置した場合，その対象になった企業には多大な被害が発生するため，風評が広がる前にその存在を認識し，適切な対策を実施する必要がある。

しかし，インターネット上には多数の掲示板サイトが存在しており，その中で日々多くの記事が発信されているため，すべての記事を詳細に確認することはできない。そこで，膨大な記事の中から問題として発展しそうな注意すべきスレッド（以下，注意スレッドと言う）を，早期にかつ自動的に抽出できる技術が求められている。

東芝は，このような背景の下，掲示板サイト上の企業評判情報を分析する風評テキストマイニング技術[(1), (2)]を開発した。

ここでは，風評テキストマイニング技術の概要とこの技術による抽出結果の妥当性について述べる。

2　風評テキストマイニング技術の概要

ここでは，風評テキストマイニング技術が対象とする掲示板サイトの構成及び，このマイニング技術の概要とそれを構成する各処理について述べる。この技術は，現在，英語記事を対象としており，日本語記事を対象とするには，日本語記事向けの分析知識を用意する必要がある。

2.1　対象掲示板サイト

掲示板サイトと一口にいっても，その形態には様々なものが

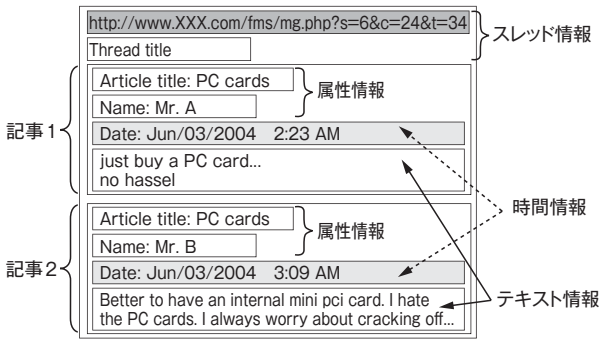


図1. 分析対象スレッドの例 — 風評テキストマイニング技術による分析では，スレッド情報，属性情報，時間情報，及びテキスト情報から成るスレッドが入力として与えられる。

Example of target thread

存在する。実際に複数の掲示板サイトを分析したところ，典型的なものは，**図1**に示すような情報から成るスレッドを格納している。そこで，図1のスレッドが格納された掲示板サイトを分析の対象とした。

図1の場合，掲示板サイトは，URL（Uniform Resource Locator：Webページのインターネット上の住所）やスレッドタイトルなどのスレッド情報と複数の記事から構成されている。また，各記事は時間情報，URLやスレッドタイトルなどの属性情報，及びテキスト情報から構成されている。

2.2 処理の流れ

風評テキストマイニング技術は，対象掲示板サイトからスレッドを収集し，そのスレッドを構成する記事の集合を取り出して入力とする。この記事の集合に対して，三つの抽出処理を順次適用して分析し，注意スレッドと，その特徴的な表現を抽出する。このマイニング技術では，あらかじめ設定されている知識を利用することでこれらの分析を行っており，その結果をデータベース（DB）に格納している。利用者は，これらの分析結果の一つとして，注意する度合いが高い順にランキングされた注意スレッドを参照することができる。また，これらの分析結果とコールセンターに蓄積されているそのほかの情報を参照することで，この注意スレッドに対する対策の必要性を判断することができる。

風評テキストマイニング技術の概要を**図2**に示し，各処理と分析知識の概要を以下に述べる。

2.2.1 ラベル抽出　風評テキストマイニング技術では，会社名，機器名，不満といった記事を特徴付ける内容が記載されている場合に，その記事を特徴付けるラベルとして，それらの内容をラベル抽出処理を用いて抽出する。今回開発した技術では，分類モデル及びラベル関連表現辞書それぞれに基づく2種類の方法によりラベルの抽出を行っている。ただし，分類モデルは，手動でラベル付けされたテキスト（訓練データ）に，機械学習法の一つであるSVM（Support Vector Machine）を適用することで学習される。この分類モデルには，テキストとラベルの有無の関係が記述されている。また，ラベル関連表現辞書には，ラベルに関連すると考えられる単語やフレーズが登録されている。以下に，各抽出法の詳細を述べる。

分類モデルに基づいたラベルの抽出では，記事の中から抽出したテキスト情報に対して，自然言語処理技術の一つである形態素解析を行う。この解析では，語尾変化など単語の変化を吸収した文字列（語幹）を抽出するとともに，各単語に対応する品詞を決定する。

このようにして解析されたテキスト情報に対して，このテキストを特徴付ける語幹が含まれているかどうかを判定し，それに基づいた特徴ベクトルを生成する。ただし，テキストを特徴付ける語幹は，分類モデルの学習時に，訓練データを構成するテキスト情報に含まれる頻度に応じて決定されている。各記事のテキスト情報から，特徴ベクトルが生成されるようすを**図3**に示す。このベクトルを分類モデルに適用して評価することにより，この記事にラベルを付与するかどうかを決定する。

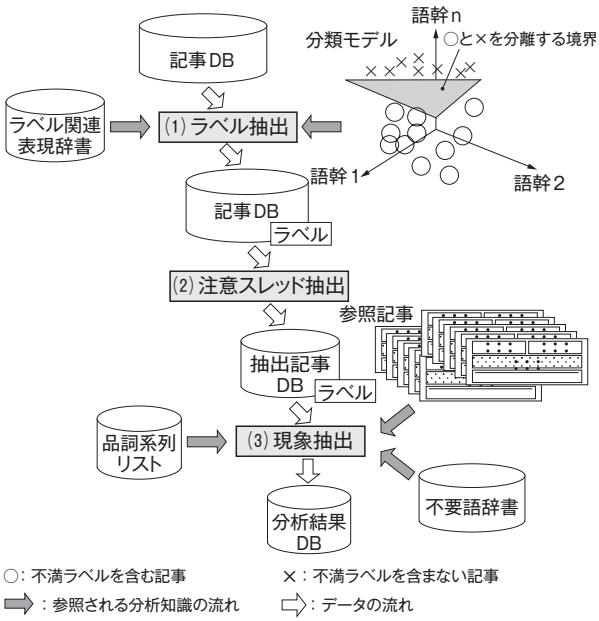
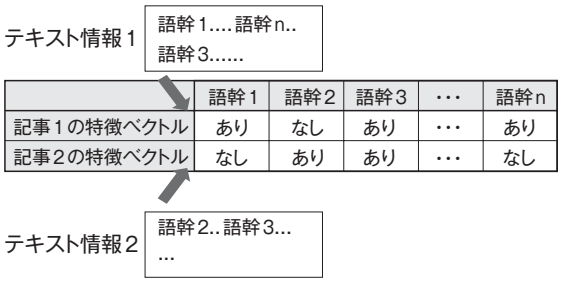


図2．開発した風評テキストマイニング技術の概要 — あらかじめ設定されている知識を利用することで，収集した記事に三つの抽出処理を順次適用して分析し，その結果をDBに格納する。

Outline of newly developed text mining technology for corporate reputation information

また，ラベル関連表現辞書に基づいたラベル抽出では，各記事におけるラベル抽出に先立ち，各ラベル関連表現に対して形態素解析を行い，対応する語幹の列を抽出する。次に，解析された記事のテキストに含まれる語幹の列とラベル関連表現の語幹の列を比較することで，この記事にラベルを付与するかどうかを決定する。

ただし，語幹の列の比較では，ラベル関連表現に対して用



	語幹1	語幹2	語幹3	・・・	語幹n
記事1の特徴ベクトル	あり	なし	あり	・・・	あり
記事2の特徴ベクトル	なし	あり	あり	・・・	なし

図3．特徴ベクトル生成のようす — 記事を構成するテキスト情報が形態素解析されて，特徴ベクトルが生成される。

Process of generating characteristic vectors

意された，語幹の列に対する3種類の制約条件，すなわち連続，系列，及び存在を利用して判定している。ここで，連続は，ラベル関連表現に対応する各語幹が形態素解析されたテキスト内に連続して出現することを表し，系列は，ラベル関連表現に対応する各語幹が解析されたテキスト内にその出現順序を守って出現することを表し，また，存在は，ラベル関連表現に対応する各語幹が解析されたテキスト内に出現することだけを表している。このような簡便な制約条件だけを利用することで，ラベル関連表現を柔軟に設定することができ，また，自然言語処理技術に関して特段の知識を持たない利用者でも，ラベル関連表現辞書を改訂することができる。

以上に述べた二つのラベル抽出処理の少なくとも一つの処理でラベルが抽出された場合に，対応する記事にラベルを付与する。このような判定を行うことで，ラベル抽出におけるラベルの見逃しを極力回避することができる。

2.2.2　注意スレッド抽出　注意スレッド抽出処理では，ラベル抽出処理で付与した会社ラベル及び不満ラベルを基準にして，多数のスレッドの中から利用者にとっての注意スレッドを抽出する。

この注意スレッド抽出では，始めに各スレッドに対して，話題の中心となっている会社を判定する。すなわち，スレッドに出現する頻度が最大の会社を，話題の中心となっている会社の候補と判定する。

次に，このスレッドやその中の会社ラベルを含む記事集合において，これらの会社ラベルが出現する割合を評価する。この評価により，複数の会社が話題として取り上げられているスレッドを無視することができるとともに，話題の会社が一つに特定されるスレッドだけを抽出することができる。このようにして抽出された会社が，あらかじめ利用者によって指定されている会社と一致する場合に，それらのスレッドを注意スレッドの候補として抽出する。

これら注意スレッドの候補に対して不満の高まりを評価することで，注意スレッドとして抽出するかどうかを決定する。具体的には，不満ラベルを割り当てられている記事の件数（不満件数）を算出し，その件数がしきい値以上となる注意スレッドの候補を注意スレッドとして抽出する。最終的には，抽出された注意スレッドを不満件数の多い順にランキングすることで，注意する度合いが高い順に出力できる。

2.2.3　現象抽出　現象抽出処理では，注意スレッドにおいて頻出するが，一般のスレッドにおいては頻出しない表現を，それらの注意スレッドにおいて話題となっている問題の原因（以下，現象と言う）として抽出する。具体的には，**図4**に示すように，参照記事，品詞系列リスト，不要語辞書といった分析知識を，現象候補抽出処理，差分解析処理，及び不要語削除処理で利用することで，各注意スレッドから現象を抽出する。ここで，参照記事は，一般のスレッドから収集された記事を格納したものである。また，品詞系列リストは，現象の候補となりうる特徴的な品詞の列を格納したものである。不要語辞書は，明らかに現象とは考えられない表現を格納したものである。以下に，図4の各処理の概要を述べる。

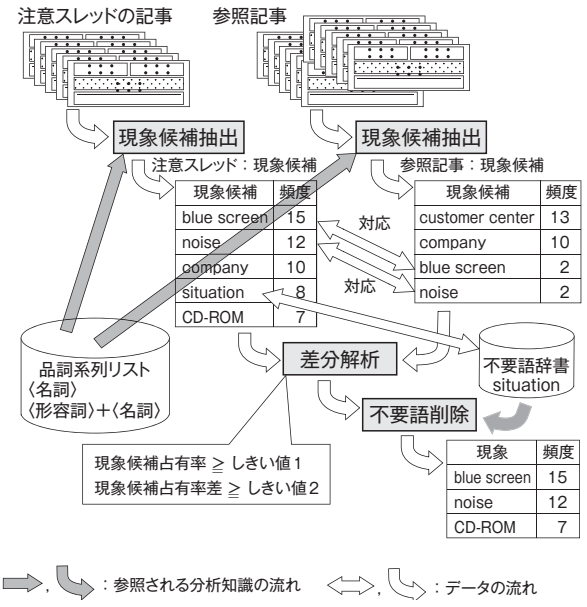


図4．現象抽出処理の概要 — 注意スレッドに対して，現象候補抽出，差分解析，及び不要語削除の各処理を実施し，現象を抽出する。

Outline of process of extracting relevant expressions

現象候補抽出処理は，系列リストで指定されている品詞列を持つ語幹の列を，注意スレッド及び参照記事の双方から抽出するとともに，その品詞の列の頻度を算出する。

次に，注意スレッドから抽出された現象候補に対して，その頻度と参照記事における頻度の比較から，差分解析処理が注意スレッドにおける現象候補の絞込みを行う。具体的には，これらの現象候補に対して，この注意スレッドにおいて出現する割合である現象候補占有率と，参照記事における現象候補占有率との差である現象候補占有率差を算出する。これら二つの値があらかじめ設定されているしきい値以上となる現象候補だけを，この注意スレッドにおける現象候補として抽出する。この差分解析を実施することで，注意スレッドにおいて頻出するが，参照記事においては頻出しない現象候補を抽出することができる。

更に，不要語削除処理によって，不要語辞書に格納されていない表現だけを現象として抽出する。この不要語辞書に格納される不要語は，抽出された現象を利用者が参照し，現象とは考えられない表現に遭遇した場合に適宜追加することができる。このため，スレッドの分析が繰り返されるのにしたがって，現象とは考えられない表現は抽出されなくなると考えられる。

3　妥当性と有効性の評価

この風評テキストマイニング技術では，スレッドはラベルによって特徴付けられており，そのラベルに基づいて注意スレッドや現象の抽出が行われている。このため，ラベルは最重要な要素であり，その抽出結果の妥当性を評価することが必要である。この技術においては，スレッドを特徴付けるものとして，会社ラベル，機器ラベル，不満ラベルなどを利用している。会社ラベルや機器ラベルなどは，表現の多様性が比較的少なく，容易に妥当なラベルを抽出することができる。これに対して，不満ラベルはその表現が多様なため，妥当なラベルの抽出は容易ではなく，抽出結果の妥当性の評価対象として適切である。そこで，この技術による抽出の妥当性を，不満ラベルの抽出の妥当性に基づいて評価した結果を以下に述べる。

不満ラベルの抽出性能を評価するにあたって，英語を対象とする四つの掲示板サイトから収集した各200件，計800件の記事を利用した。これらの記事の集合を開発した技術に入力として与えることにより，特定の企業に対する不満が記載されているかどうかを判定した。一方，長年このような分析業務に携わっている利用者が，実際にこれらの記事を精読し，特定企業に関する不満を示した内容が記載されているかどうかを判定した。そして，開発した技術と人間系（利用者）による判定結果の一致度合いを見ることで，この技術の抽出性能の妥当性を評価した。

表1は，これら評価結果として，この技術と利用者による判定結果の差を示している。表1からわかるように，開発した技術による判定が利用者によるそれと一致した記事の件数が776件あり，この技術の正解率は97.0％であった。また，再現率は82.3％，適合率は86.7％であった。なお，再現率は不満を含む記事がどのくらい抽出できたかを，適合率は抽出した記事がどのくらい正確に不満を含む記事であったかを示す指標である。

これらの指標のうち，特に再現率は，真の注意スレッドを見逃さない確率を評価するうえで重要な指標である。ここでは，数値的な目標を設定するため，真の注意スレッドには不満を含む記事が少なくとも三つ存在することを仮定している。また，そのすべての記事の抽出に失敗した場合に，真の注意スレッドを見逃すことになると仮定している。これらの仮定の下で，真の注意スレッドを見逃す確率を1.0％以下に抑えることを目標に掲げており，そのためには，再現率を80.0％以上にする必要がある。ここで，不満を含む記事の数を3とすることで，$(1-0.8)^3 \leqq 0.01$となり，目標設定が正当と考えられる。

今回の評価実験における再現率はこの目標値を超えており，利用者にとって妥当な抽出性能を実現しているといえる。また，利用者による直感的な評価においても，開発技術によって妥当な抽出が行われているとのコメントを得ており，この技術の有効性を確認することができた。

表1．開発技術による不満ラベル抽出性能の評価結果

Results of evaluating complaint label extraction performance using newly developed text mining technology

（単位：件）

		開発技術	
		不満	非不満
利用者	不満	65	14
	非不満	10	711

●総記事数：4掲示板サイト×200件＝800件
●開発した技術と利用者の判定が一致した記事数　65件＋711件＝776件
●開発した技術による正解率：(65件＋711件)/800件＝97.0％
●開発した技術による再現率：65件/(65件＋14件)＝82.3％
●開発した技術による適合率：65件/(65件＋10件)＝86.7％

4　あとがき

掲示板サイトにおける多数の記事を分析する手法として風評テキストマイニング技術を開発した。また，このような技術が有効に機能するために重要なラベルの抽出性能を評価し，開発技術が妥当な性能を備えていることを確認できた。

開発した風評テキストマイニング技術は，記事を分析するという基本機能だけでなく，掲示板サイトからの記事の収集や分析結果の提示といった機能を加えることにより，リスクサーチシステム(3)として実用化されている。

インターネットに投稿されるデータは今後もますます増えるとともに，その中でいっそう多様な意見が顧客から発信されるようになると予想される。一方，企業間競争はその激しさを増しており，顧客満足度を高めていくことがより重要になると考えられる。このため，インターネット上に埋もれている企業の評判情報を分析する風評テキストマイニング技術を，今後も更に発展させていく。

文　献

(1) Sakurai, S., et al. Discovery of Important Threads from Bulletin Board Sites. Intl. J. of Information Technology and Intelligent Computing. 1, 1, 2006, p.217－228.
(2) 櫻井茂明，ほか．掲示板サイト分析における重要議論抽出と特徴表現抽出．知能と情報．19, 1, 2007, p.13－21.
(3) 安齋学徳，ほか．掲示板サイトの風評情報分析システム．東芝レビュー．62, 10, 2007, p.54－57.

櫻井　茂明　SAKURAI Shigeaki，Ph.D.
研究開発センター　システム技術ラボラトリー研究主務，博士（工学）。データマイニング技術に関する研究・開発に従事。日本知能情報ファジィ学会会員，技術士（情報工学）。
System Engineering Lab.