掲示板サイトの風評情報分析システム

System for Analysis of Rumor Information on Bulletin Board Sites

安齋　学徳　　櫻井　茂明
■ ANZAI Takanori　■ SAKURAI Shigeaki

掲示板サイトには製品や企業に対する多くの風評情報が記載されており，その内容を分析することにより，製品や会社に対する要望や不満などの傾向を知ることができる。特に近年，掲示板サイトの社会的影響力は増加傾向にあり，掲示板サイトからユーザーの意見が盛り上がり，集団訴訟などに発展することもある。このため，掲示板サイトを監視して製品や会社に対する風評情報を常にチェックし，リスクに備えることが重要になってきている。

東芝は，このようなニーズに応えるため，掲示板サイトに記載されている内容を自動的に分析し，特定の製品や会社に対する風評情報をレポートするシステムを開発した。

Many rumors and speculations concerning enterprises and their products appear on the sites of bulletin board systems (BBS). It is possible to gain an understanding of the trends in users' requirements and points of dissatisfaction by analyzing such information. With their recent increase in social influence, BBS can also be the source of class actions and other movements. It is therefore important to observe BBS sites at all times and check the rumor information that they contain in order to be prepared for future risks.

To meet these requirements, Toshiba has developed a rumor information analysis system that analyzes rumors related to specific companies and products and reports the analysis results.

1　まえがき

掲示板サイトの風評情報を分析する場合，調査対象となるサイトやキーワードを指定し，対象としたサイトからキーワードに関連する記事を収集し，その内容を確認するのが一般的である。

分析精度を高めるには，分析したい分野に関連するできるだけ多くのサイトから風評情報を収集して分析する必要がある。しかし，そうすると確認しなければならない記事数は月に数万件に上り，収集した記事を一つ一つ確認するのに多くの時間を要してしまう。このため，専門業者に分析を依頼することも多く行われているが，依頼コストは対象サイト数やキーワード数に比例して増加する。

そこで，東芝は掲示板サイトの風評情報の分析を効率よく，かつ精度よく行うことを目的に，掲示板サイトの風評情報を分析するシステムを開発した。

ここでは，そのシステムの概要と特長となる機能について述べる。

2　システムの概要

風評情報分析システム（**図1**）は，掲示板サイトに投稿された記事を自動的に分析し，製品や会社に対する不満の有無をわかりやすくレポートするシステムである。掲示板サイトから収集した記事を投稿者名，投稿日，記事タイトル，記事本文に分解し，記事タイトルや記事本文内に，利用者があらかじめ設定した製品や会社に対する不満の記述がないかを分析する。分析した結果，不満件数を月別や週別に投稿日を使ってカウントし，不満件数の時系列変化を出力する。このとき不満の有無は，単純な不満を意味する単語が記事内にある，なしだけで判定するのではなく，文の構造などを分析して判定する。

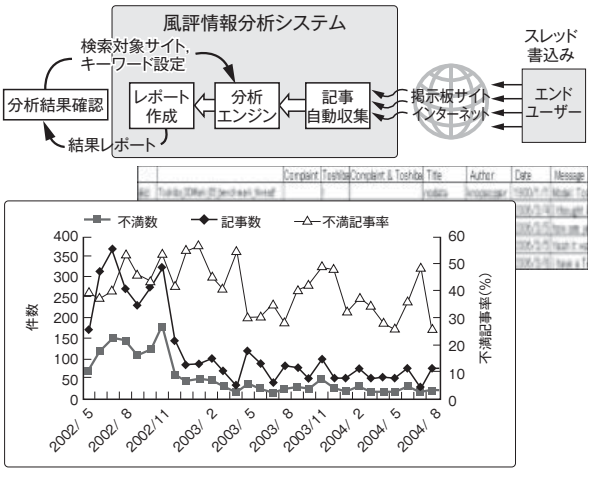

図1. 風評情報分析システム — 掲示板サイト内の記事を自動的に収集，分析し，分析結果を視覚的にレポートする。
Outline of rumor information analysis system

このシステムは次の機能から成り，記事収集から分析，レポート作成までをすべて自動的に行う。

(1) 掲示板サイトから記事を収集する機能
(2) 収集した記事内に不満の記述があるかを分析する機能（分析エンジン）
(3) 分析した結果からレポートを作成する機能

このため，システムをバックグランドで実行させておけば，製品や会社への不満件数の推移及び不満の具体的内容を定期的に確認することができる。

3 システムの特長

システムの主な特長は，次のとおりである。

3.1 記事自動収集機能

記事自動収集機能は，利用者が指定したサイトの下に存在するすべての掲示板サイトを検出し，そこに投稿されたすべての記事を収集する機能である。サイトの下のすべてを対象にすることで，サイトの下に新規に立ち上がった掲示板やスレッド(注1)も収集対象になる。記事自動収集機能の概要を以下に述べる。

3.1.1 記事検出方法 Webサイトは通常，サイト，掲示板，スレッド，記事のツリー構造になっている。記事自動収集機能は，まずサイト内のどこに掲示板やスレッドがあるかを検出する必要がある。しかし，サイトのフォーマットは統一されていないため，掲示板の存在する場所を容易に特定することができない。

そこでこの処理では，記事収集対象サイトのHTML (HyperText Markup Language) 文書をXML (eXtensible Markup Language) 化し，XQueryを利用してXML文書から必要な情報を収集する方法を採用した。XQueryはXML文章に対して，様々な問合せが行える言語である。この方法だとあらかじめXQueryをサイトごとに準備しておく必要があるが，XQueryをプログラム外部に置くことで，基本プログラムを変更することなしに，任意のサイトから必要な情報を収集できるようになる。この処理では，XQueryを使ってXML文書から“スレッドタイトル，投稿者名，投稿日，記事タイトル，記事本文”を識別して収集し，3.2節で述べる分析エンジンの入力データ（収集記事データベース（DB））としている。

3.1.2 自動認証 認証が必要なサイトに対しては，自動認証を行う機能を作成した。利用者が一度，ユーザー登録などサイトを参照するために必要な手続きをして参照権限が得られると，その後は，プログラムが認証処理を自動的に行うため，効率的な情報収集を行うことができる。

3.1.3 記事収集対象外の指定 利用者が指定したサイトの下には，利用者にとってまったく関係ない分野の掲示板やスレッドが存在することがある。このため，明らかに関係のない掲示板やスレッドを記事収集の対象外にする機能を作成した。

記事収集対象外設定用GUI (Graphical User Interface) に，サイト内の全掲示板名とスレッド名の一覧をツリー構造で表示する。利用者は，表示されたリストから，記事収集対象外にする掲示板名又はスレッド名をマウスで選択する（**図2**）。掲示板名を指定すると，その掲示板内のすべてのサブ掲示板とスレッドが記事収集の対象外になる。

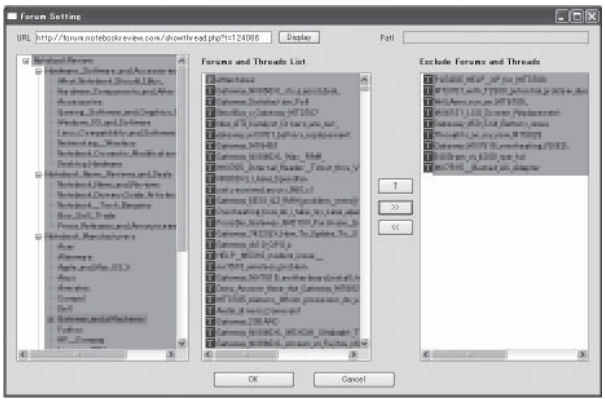

図2. 記事収集対象外指定機能 — サイト内の掲示板とスレッドの一覧がツリー構造で表示され，そこから除外する掲示板名又はスレッド名を選択指定できる。
Setting of threads for exclusion

3.1.4 記事差分収集 利用者が指定したサイトの下のすべての情報を毎回収集するのは効率的でないため，過去に収集した情報を再収集しないよう，最後に記事を収集した日以降に投稿された記事だけを収集する機能を開発した。

3.2 分析エンジン

分析エンジンは，収集記事DBに格納されているスレッド単位の記事を入力とし，指定された会社に関する不満が多数記述されているスレッド（以下，注意スレッドと記す）を順位付けして，注意スレッドを代表する表現とともに分析結果DBに出力する。この分析エンジンは，**図3**に示す分類モデル及び，イベント関連単語，抽出品詞系列リスト，参照記事，不要語辞書と呼ばれる背景知識を利用しつつ，イベント抽出及び，注意スレッド抽出，現象抽出と呼ばれる三つの抽出処理を順次実施することで，記事の分析を行っている[1), (2)]。各抽出処理の概要について以下に述べる。

3.2.1 イベント抽出 イベントの抽出は，分類モデルとイベント関連単語の二つの背景知識を利用して行っている。ここで，イベントとは，注意スレッドかどうかを判断するうえで必要な記事に記述されている主体や，行動，感情などを代表する

(注1) ある特定の話題に関する投稿の集まり。

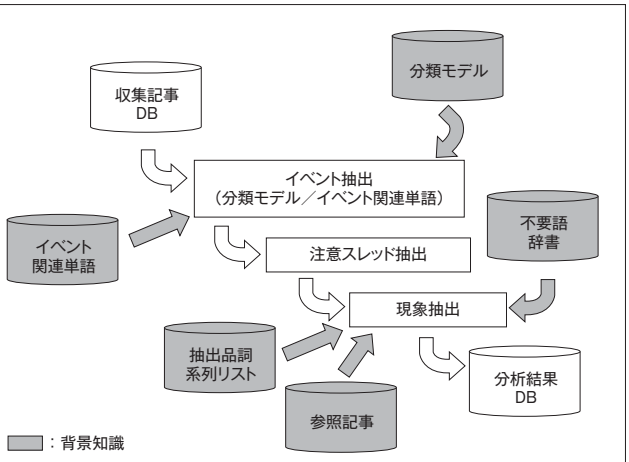

図3. 分析エンジンの概要 ― 5種類の背景知識を利用した3種類の抽出処理に基づき，収集記事を分析する。

Outline of risk search analyzer

ものである。このシステムでは，会社名と不満の有無がイベントとして定義されている（以下それぞれを，会社イベント，不満イベントと記す）。イベント抽出では，各背景知識に基づいた方法によって個別にイベントを抽出しており，抽出された結果を統合することで各記事に付与すべきイベントを決定している。

分類モデルに基づいたイベント抽出では，機械学習手法の一つであるSVM（Support Vector Machine）と呼ばれる手法を利用することで，あらかじめ分類モデルを学習している。この分類モデルは，多数の重要な単語の語幹で構成される空間上にある，特定のイベントの有無を識別する超平面として与えられる。このシステムでは，約10,000件の記事と各記事に付与されたイベントの組から，分類モデルを学習している。分類モデルに基づいたイベント抽出では，各記事のタイトル及び本文部分に対して，自然言語処理技術の一つである形態素解析を実施することで，タイトル及び本文部分から単語の語幹と品詞から成るデータ（以下，形態素解析結果データと記す）を記事ごとに生成する。また，分類モデルを構成する重要な単語の語幹をこの形態素解析結果データが含んでいるかどうかを識別することで，該当する語幹の有無によって構成される属性値ベクトルを記事ごとに生成する。この属性値ベクトルをイベントに対応する分類モデルに適用することで，対象とする記事に該当するイベントの有無を識別し，イベントがあると識別された場合には，その記事に該当するイベントを付与する。

一方，イベント関連単語に基づいたイベント抽出では，イベントを表す複数の表現を記述した辞書を設定している。例えば，英語を対象とする場合，“displeasure”，“frustration”，“not satisfy”などの表現が，不満イベントに対応するイベント関連単語として辞書に登録されている。イベント関連単語に基づいた抽出では，各イベント関連単語にあらかじめ形態素解析を適用しておくことで，対応する語幹と品詞を決定している。また，形態素解析結果データにイベント関連単語の形態素解析結果が含まれているかどうかを比較し，イベント関連単語が含まれていると判定された場合は，その記事にイベント関連単語に対応するイベントを付与する。

このシステムでは，イベントの抽出精度と抽出時間を勘案し，不満イベントに対してだけ二つの抽出方法を利用し，ほかのイベントにはイベント関連単語に基づいた抽出方法だけを利用している。

3.2.2 注意スレッド抽出　注意スレッドの抽出は，イベント抽出によって各記事に付与されたイベントを参照することで，多数のスレッドの中から注意スレッドを抽出し，その順位を決定する。このシステムでは，対象会社判定，不満件数判定，スレッド順位付けといった三つのサブ処理を順次実施することで，順位付きの注意スレッドを抽出している。

はじめに，対象会社判定は，スレッドを構成する記事に付与されている会社イベントを会社名ごとに収集し，その出現頻度を算出する。また，この出現頻度の割合に基づいて，スレッドで主に議論されている会社イベントを特定する。特定された会社イベントが，あらかじめ指定されている会社イベントと一致する場合には，このスレッドを注意スレッドの候補とする。

次に，不満件数判定は，注意スレッドの候補を構成する記事に付与されている不満イベントを収集し，その出現頻度を算出する。また，この出現頻度があらかじめ指定されている最小不満件数以上になるかどうかを判定し，不満件数以上となる場合に，この候補を注意スレッドとして判定する。

最終的には，スレッド順位付けは，抽出された注意スレッドに対して，不満イベントの出現頻度を参照することにより，不満イベントの出現頻度の高い順に注意スレッドを出力する。

3.2.3 現象抽出　現象抽出は，抽出品詞系列リスト，参照記事，及び不要語辞書の三つの背景知識を利用することで，注意スレッドを特徴付ける表現として，注意スレッドから現象を抽出している。この現象抽出は，表現抽出や，差分解析，不要語削除といった三つのサブ処理を順次実施することで実現されている。

はじめに，表現抽出は，スレッドの内容にとって重要な品詞の組合せである抽出品詞系列リストを，形態素解析結果データに適用することで，登録されている抽出品詞系列に一致する単語列を抽出し，その出現頻度を算出する。

次に，差分解析は，抽出品詞系列リストを特定のビジネス領域について記述されている参照記事に適用することで，参照記事に含まれる単語列とその頻度をあらかじめ算出する。このシステムでは，約12,000件の記事が参照記事として格納されている。この差分解析は，注意スレッドから抽出された単語列の出現頻度と，その単語列の参照記事での出現頻度とを比較することで，注意スレッドでは頻出しても参照記事では頻出しない単語列を，その注意スレッドに対する現象の候補として抽

出する。

最終的には，不要語削除は，現象とは考えらない表現を登録している不要語辞書にその候補が格納されているかどうかを判定し，格納されていない場合には現象として出力する。ただし，不要語辞書は，利用者により適宜表現の追加が可能となっており，利用者の意図を反映した現象だけを抽出することができる。

3.3 レポート機能

レポート機能は，分析エンジンが分析した結果を視覚的にわかりやすく出力する機能である。レポート機能の概要について以下に述べる。

3.3.1 記事件数・不満記事件数の時系列変化出力

記事件数とは，一定期間内に指定サイト内の掲示板に投稿された記事の総件数であり，不満記事件数とは，利用者が指定した製品や会社に対して不満を述べている記事の件数である。記事の総件数と不満記事件数の時系列変化をグラフに出力し，不満記事件数の急増などを視覚的に確認できるようにした（**図4**）。

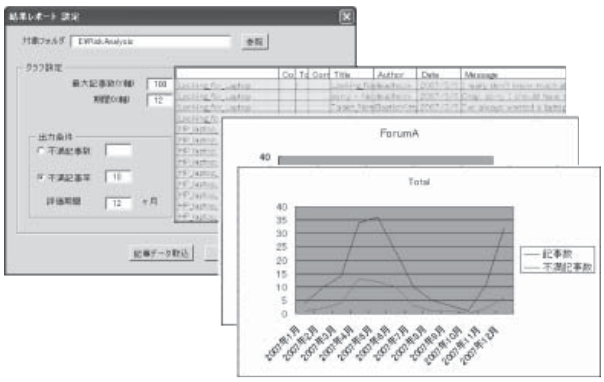

図4. レポートの例 — 記事件数と不満記事件数の月別グラフ，及び不満記事本文の一覧などがボタン一つで出力できる。
Example of report

3.3.2 記事内に多く登場する現象名のランキング出力

記事内に多く登場する現象名は，世の中が注目している現象である可能性が高い。そこで，注目されている現象が確認できるように，記事内に登場する回数の多い現象名をランキング出力し，登場件数の時系列変化もグラフ出力できるようにした。

3.3.3 異常値メール通知　記事件数や不満記事件数が任意に決めた件数以上あり，かつ次の条件と一致した場合，メールで不満内容を利用者に通知する機能を開発した。

(1) ある期間内の記事件数又は不満記事件数の増加量が任意に設定された値を超えた場合
(2) 記事件数又は不満記事件数が任意に決めた期間以上連続で増加している場合

3.3.4 記事内容出力　掲示板サイトから収集した生の記事データを目視確認する機能を開発した。このとき，必要な記事だけを容易に確認できるように，次の条件でフィルタリングできるようにした。各条件は同時に複数指定でき，指定した条件すべてを満たす記事が確認できる。

(1) 不満記述が含まれる記事
(2) 利用者が指定した製品名や会社名が含まれる記事
(3) 利用者が指定した現象名が含まれる記事
(4) 利用者が指定した期間内に投稿された記事
(5) 利用者が指定した掲示板サイトに投稿された記事

4 あとがき

ここで述べた風評情報分析システムを利用することにより，掲示板サイトに投稿された記事の中から，利用者が指定した製品や会社に対する不満記事が自動的に抽出できるようになり，不満の分析が効率よく行えるようになった。

今後は，分析エンジンを更に改善して不満認識の精度を高めるとともに，あらかじめ対象となる掲示板サイトを指定せずに分析できる方法を検討していく。

文 献

(1) Sakurai, S., et al. Discovery of Important Threads form Bulletin Board Sites. Int. 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.

安齋　学徳　ANZAI Takanori
PC&ネットワーク社 PC開発センター 設計プロセス開発センター第一担当主務。
社内情報システム開発に従事。
PC Development Center

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，工博。機械学習技術に関する研究・開発に従事。日本知能情報ファジィ学会監事。技術士（情報工学）。
System Engineering Lab.