インターネット情報監視システムの試作

永井 明人　増塩 智宏　高山 泰博　鈴木 克志

三菱電機株式会社　情報技術総合研究所

1. はじめに

インターネットでは一般からの情報発信が盛んになり、企業や製品に関する消費者の生の声(風評)が広く公開されるようになった。そこで、これらの大量の風評からクレームを抽出して、迅速なクレーム対応を実現する要求が企業において急速に高まっている。こうした要求を背景として、Web 上に広がる企業や製品のクレーム情報を抽出して監視するインターネット情報監視システムを試作した。本稿では、この試作システムの概要を述べる。

2. インターネット情報監視の課題

大量の Web 文書を対象とした情報監視の業務では、以下が課題となる。

(1) 一般の全文検索エンジンでは、検索結果として取得できる URL 数に上限があり、大量に収集できない。また、索引付けに時間がかかるため、最新情報の検索が困難である。

(2) 大量文書から、クレーム文書を人手で判断して抽出するのが困難である。また、既存の全文検索エンジンや風評配信サービスでは、クレーム文書を検索するためのキーワードの設定が困難である。

(3) 急速に広がりつつあるクレームを迅速に把握することが困難である。

そこで、本システムでは、上記課題に対して以下の解決手段を実現した。

(1) 検索結果として取得できる URL 数の上限を超えて収集するために、時分割収集を行なう。さらに、最新情報である掲示板などの特定 URL 監視を行なう。

(2) 単語共起に基づくクレーム抽出技術[1][2][3]により精密なパタン照合を行なって、クレーム文書を自動抽出する。

(3) クレーム文書のマクロな時系列分析を行なうトレンド分析により、危機の予兆を迅速に検知し、クレーム対応を支援する。

“Prototyping an internet watching system”
NAGAI Akito, MASUSHIO Tomohiro, TAKAYAMA Yasuhiro, SUZUKI Katsushi
Information Technology R&D Center
Mitsubishi Electric Corporation

3. システム構成

本システムは図 1に示すように、Web 文書収集部、情報抽出(クレーム抽出)部、トレンド分析部の三つの処理から構成される。

処理の流れとしては、オペレータが調査対象に関する初期設定として、例えば自社の企業名や、調査対象を表わす簡単なキーワード(製品のカテゴリ名)などをプロファイルデータとして設定する。システムは、プロファイルデータに基づき、定期的にインターネットから文書を収集し、収集した文書に対してクレーム抽出を行なう。さらに、クレームを判定された文書集合に対し、トレンド分析を行ない、クレーム出現傾向を視覚化表示する。

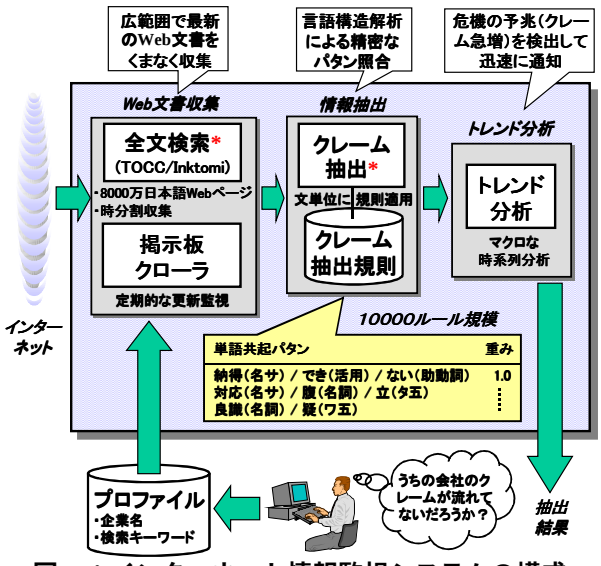

図 1：インターネット情報監視システムの構成

4. Web 文書収集部

文書の収集処理は、図 2に示すように全文検索部、ダウンロード部、および掲示板クローラ部からなる。

全文検索では、プロファイルデータ中の企業名と検索キーワードを入力として、調査対象に関する Web 文書の URL リストを取得する。この際、時分割収集のために、全文検索エンジン TOCC[4]の機能を用いて検索し、取得した URL リストをダウンロード部へ渡す。

ダウンロード部では、URL リストの各 URL からテキスト情報を取得し、また、掲示板クローラ部では、特定 URL にある掲示板に対してクローリングを行ない、各発言ごとにテキスト情報を取得する。

上記のテキスト情報はディスク上に格納し、また、収集日時、Web 文書の更新日時、掲示板発言の発言日時といった書誌情報は、URL・文書管理 DB へ記録して管理する。

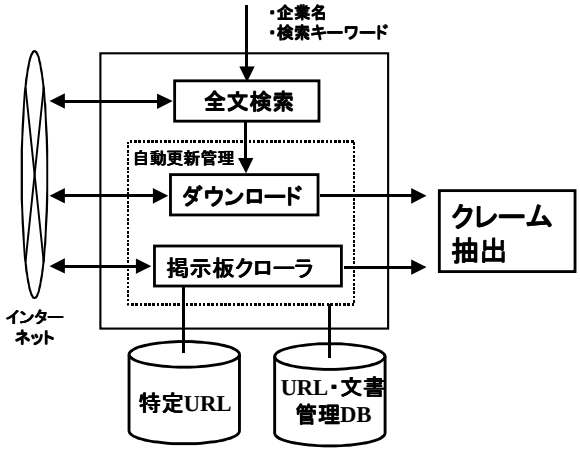

図 2：Web 文書収集部

5. クレーム抽出部

収集した Web 文書に対して、文献[1][2][3]の方式に基づくクレーム抽出を行なう。本方式は、意図(クレーム)を表現する一般的な特徴表現を、複数の単語の共起パタンとして規則化し、意図抽出を行なうアプローチであり、図 1に示すような単語共起パタンを 1 万ルール規模で適用している。

クレーム抽出処理では、入力された文書を文単位の解析単位に分割し、形態素解析の後、クレーム抽出規則を参照して、解析単位中の形態素列と単語共起パタンとの照合を行なう。単語共起パタンが解析単位の形態素列に存在すれば、文書に対するクレーム度スコアにクレーム抽出規則の重みを加算・正規化し、クレーム度スコアが閾値を越えた場合に、文書をクレーム文書と判定して、抽出表現と共に出力する(図 3)。

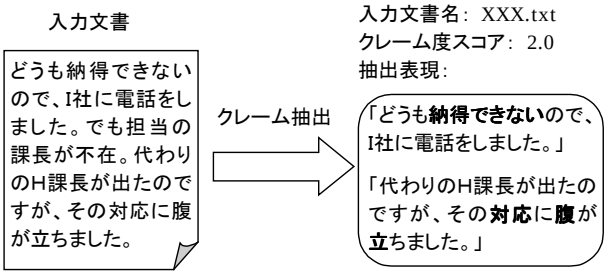

図 3：クレーム抽出結果の例

さらに、抽出表現の近傍に存在する企業名を抽出し、URL・文書管理 DB へクレーム抽出結果とともに格納する。

6. トレンド分析部

トレンド分析部は、抽出したクレームの出現傾向を時系列でマクロに把握するために、URL・文書管理 DB に格納されたクレームのスコアの推移をグラフとして表示する。この機能により、Web 上にクレームが急増し始めたことを検知し、迅速なクレーム対応を支援する。図 4は、製品 X に関するクレーム急増の実際の分析例である。製品 X が発売されて不具合が発覚し、クレームが急増していることが分かる。トレンド分析の機能により、本事件の新聞報道日以前に危機の予兆を把握することができるようになる。

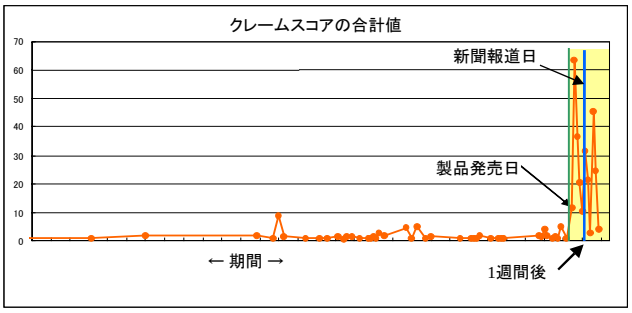

図 4：クレーム急増検知の例

7. おわりに

クレーム抽出技術を適用した応用システムとして、インターネット上の風評情報を監視するシステムを試作した。今後は、試作システムの実験評価を実施し、応用システムとしての業務効果を、定量データとして明確化していく予定である。また、業務支援のために有効な機能も検討していく。

[参考文献]

[1] 永井, 他 “CRM における顧客メール分析手法の検討,” 情報処理学会 第 62 回(平成 12 年後期)全国大会 3-81, 2000.

[2] 永井, 他 “文内の単語共起照合に基づくクレーム抽出方式の性能評価,” 情報処理学会 第 64 回(平成 13 年後期)全国大会 pp. 3-17, 2002.3.

[3] 永井, 他 “単語共起照合に基づくクレーム抽出方式の改良,” FIT2002 情報科学技術フォーラム E-16, pp. 113-114, 2002.9.

[4] http://www.tocc.co.jp/search/