自 然 言 語 処 理 154－18
（２００３．３．７）

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

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。特徴は、(1) 文内の単語共起照合に基づく精密なクレーム抽出、(2) 収集したクレーム情報をマクロに時系列分析して、クレームの急増を検知するトレンド分析、(3) Web 全文検索エンジンと掲示板クローラを組合わせた、大量・最新文書の収集、である。本稿では、この試作システムの概要を述べる。

Prototyping an internet watching system

Nagai Akito , Masushio Tomohiro , Takayama Yasuhiro , Suzuki Katsushi

This paper describes an internet watching system which enables to extract consumer claims automatically from an internet. Reputation of enterprises or products latent so far is coming to appear and spread fast in an internet because everyone can send and read many messages easily in the internet. Then, it is highly required to find claims for the enterprises in order to cope with the claims quickly in terms of risk management. So we have developed and prototyped the system which is characterized by technologies of automatic claim extraction, trend analysis of claims and collection of numerous and latest documents.

三菱電機株式会社 情報技術総合研究所
Information Technology R&D Center, Mitsubishi Electric Corporation

1. はじめに

現在、EC の拡大に伴い、データウエアハウスやコールセンタへの顧客メール数が急増している。また、このような企業内文書だけでなく、インターネットにおいても、一般ユーザ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や風評情報が広く公開されるようになった。

そこで、これらの大量文書からクレーム情報を抽出して、クレームへの迅速な対応や、顧客の潜在ニーズ発掘などを実現する要求が、企業において急速に高まっている。

こうした要求を背景に、我々は、CRM における顧客メールからのクレーム抽出を目的として、文内の単語共起照合に基づくクレーム抽出方式[1]を提案し、Web 文書を対象として性能評価を行なってきた[2][3]。

さらに、本方式を Web 文書へ適用した応用システムとして、インターネット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

従来から、企業や製品などの特定の対象に関して、消費者が持つブランドイメージやニーズ、さらには風評情報等を探る市場調査が行われている。特に最近では、相次ぐ企業の不祥事を背景として、自社の危機管理とモラル向上のために、自社の風評を調査する業務も重要になりつつある。この業務では、目的の風評情報を得るための情報源として、最新で大量の口コミ情報が存在するインターネット上の Web 文書に注目が集まっている。このような Web 文書を対象とした風評調査業務では、以下が課題となる。

(1) **文書収集**：一般の全文検索エンジンは、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を要し、最新情報の検索が困難である。

(2) **クレーム抽出**：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) **マクロな傾向の把握**：インターネット上で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。

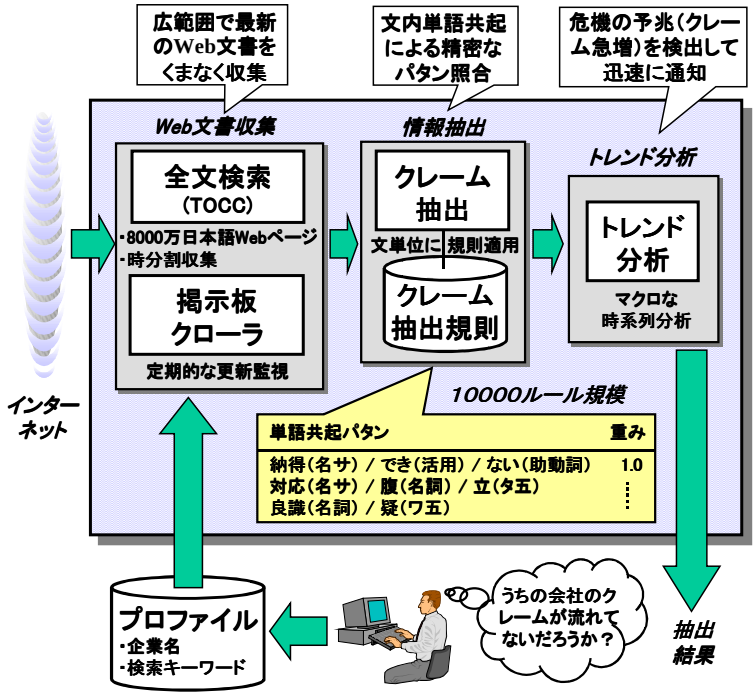

図 1:インターネット情報監視システム

そこで、本システムでは、上記課題に対して以下のアプローチにより解決を図った。

(1) **文書収集**：検索結果として取得できる URL 数の上限を超えて大量に収集するために時分割収集を行なう。さらに、最新の情報を得るために掲示板等の特定 URL を監視する。
(2) **クレーム抽出**：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。これにより、クレーム文書を検索するためのキーワードの設定が不要となる。
(3) **マクロな傾向の把握**：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。

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。

4. Web 文書収集部

Web 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。

全文検索部では、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インターネットを検索し、検索結果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 へアクセスしてテキスト情報を取得する。ここでは、大量文書の効率的な収集を行なうために、以下の機能を実装した。

- ダウンロードのマルチスレッド処理
- 各スレッドのタイムアウト時間の任意指定

また、掲示板クローラ部では、特定 URL として設定された掲示板をクローリングし、各発言ごとに分割したテキスト情報を取得する。

これら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。

上記の Web 文書収集処理は、予め設定された更新期間毎に自動実行される。例えば毎晩バッチ的に実行し、翌朝に調査担当者が抽出結果を調査するという運用を実現している。

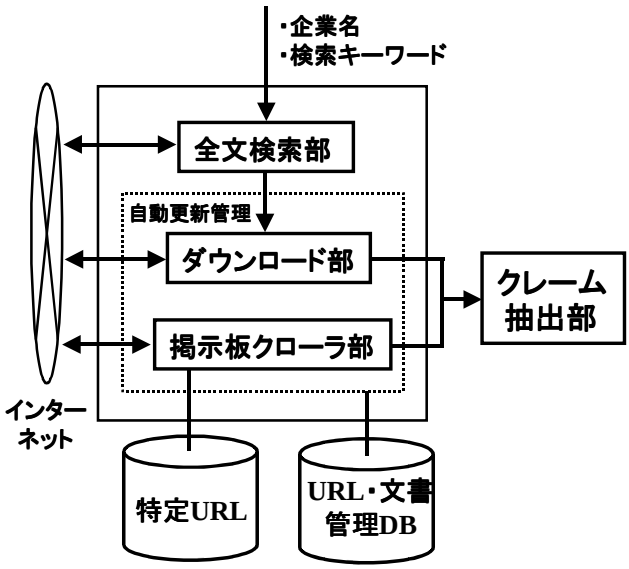

図 2:Web 文書収集部

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。

従来から、企業内文書向けに、特定の意図や意見を抽出・分類する技術として、意図認識技術[5]や、メール自動分類技術[6]などがあるが、分析対象となる業務に依存したテンプレートや辞書などの抽出知識を要し、幅広い内容を含む Web 文書への適用が困難である。

一方、インターネット上の Web 文書を対象にした風評の自動抽出技術・サービスに関して

は、製品の評判抽出[7]、風評配信サービス[8][9]などがあるが、これらは、クレームと判定するための抽出表現を単語として照合するため、複数の単語により意味を成すクレーム表現を抽出できなかった。

これらに対し、我々のアプローチでは、クレーム抽出知識を業務依存ではなく一般的な特徴表現とすることで、幅広い種々の内容を含むWeb 文書へ適用できるようにした。さらに、クレーム抽出知識として複数の単語の共起パタンを規則化して用いることにより、複数の単語により意味を成すクレーム表現を抽出できるようにした。

図 3にクレーム抽出部の構成を示す。まず、文内の単語共起照合を行なうために、入力された文書 Dを文単位の解析単位に分割する。次に、形態素解析の後、単語見出しと品詞情報を含む形態素解析結果がクレーム抽出部へ入力される。クレーム抽出では、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。

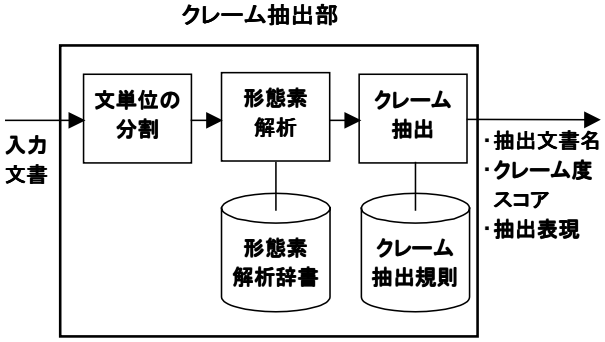

図 3:クレーム抽出部の構成

クレーム抽出規則は、表 1に示すように、単語見出しと品詞の複数の組で表現された単語共起パタンに、クレームの度合いを表わす重みを付与して定義される。単語共起パタンの照合では、各単語の順序関係が保持され、また、単語共起パタンの各単語間に許容する単語数には所定の制限を設けている。

本システムでは、このようなクレーム抽出規則を 1 万ルール規模で適用している。表 2に、実際に定義した単語共起パタンの例を示す。

表 1:クレーム抽出規則の例

No.	単語共起パタン	重み
1	納得(名サ)/でき(活用)/ない(助動詞)	1.0
2	対応(名サ)/腹(名詞)/立(タ五)	…
3	良識(名詞)/疑(ワ五)	…
…	…	…

表 2:単語共起パタンの例

＜たらい、回＞	＜常識、疑問＞
＜憤り、感じ＞	＜交渉、余地、ない＞
＜責任、逃れ＞	＜信じ、られ、ない＞
＜対応、稚拙＞	＜絶対、する、か＞
＜説明、おそまつ＞	＜ふざけ、る、な＞
：	：

これらの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 Dに対するクレーム度スコア Sに、クレーム抽出規則の重みを加算していく。文書 D全体の照合が終了した際に、クレーム度スコア Sを正規化し、所定の閾値を越えた場合に、文書 Dをクレーム文書と判定して、抽出表現と共に出力する(図 4参照)。

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。

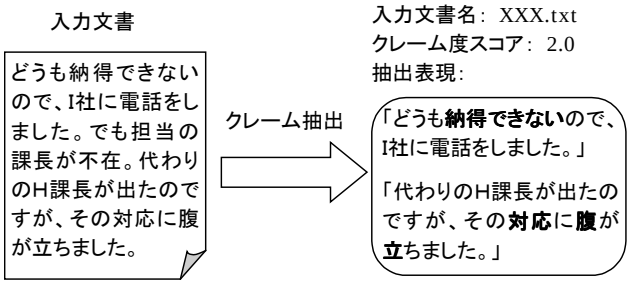

図 4：クレーム抽出結果の例

6. トレンド分析部

トレンド分析部では、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度スコアの推移を分析して、**図** 5のようにグラフとして表示することができる。

また、Web 上にクレームが急増し始めた場合には、クレーム度スコアの変動量により急増を検知して調査担当者に通知することにより、迅速なクレーム対応を支援することができる。

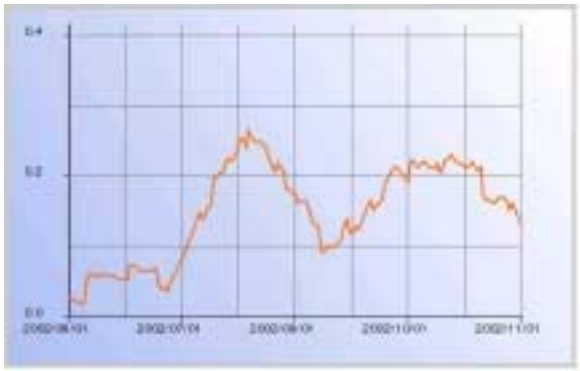

図 5 :トレンド分析の表示例

7. 評価

試作システムの各処理に関して、実際の Web 文書を用いて評価を行なった。

7.1. Web 文書収集部

ダウンロード部におけるマルチスレッド処理の効果を検証するために、スレッド数を変えた場合の通信速度を測定した(表 3)。測定環境は、CPU が Pentium III 1.2 GHz、メモリが 512 MB、通信回線は ADSL 1.5 Mbps である。実験に使用した回線の平均実効速度を測定した結果、1.2 Mbps であったため、今回使用した回線の帯域を有効に使用するのに必要なスレッド数は 20 であることが分かった。この条件下で Web 文書を収集し、100 万件規模の文書収集を実現した。

表 3:ダウンロード部の処理性能

スレッド数	処理速度 (平均)
1	203 Kbps
10	785 Kbps
20	1.1 Mbps
50	1.1 Mbps

7.2. クレーム抽出部

Web 文書に対するクレーム抽出性能を評価するために実験を行なった[2][3]。

全文検索エンジンにより Web から収集した 3215 文書を評価文書セットとした。これらに対し、表 4に示す判定基準に従ってクレーム文書と非クレーム文書とを人手で判定して、正解文書リストを作成した。クレームと判定された文書数は 97 である。なお、クレームか、非クレームかの判断に迷った場合は、非クレーム文書とした。実験で用いたクレーム抽出規則は、約 1 万ルール規模である。

図 4はクレーム抽出結果の一例であり、図中、横軸はクレーム判定のスコア閾値を表わし、縦軸は各閾値における再現率と適合率を表わす。これより、風評抽出業務で重要と考えられる再現率が高い領域において、適合率に改善の余地が残されているものの、一般的なクレーム表現の 9 割以上をカバーしていることが分かった。

表 4:クレーム/非クレームの判定基準

クレーム	一般的なクレーム表現。 製品の不具合情報に関する表現。 口論相手への明確なクレーム表現。
非クレーム	一般的なクレーム表現を含まない文書。 製品の不具合情報を含まない文書。 クレームの相談窓口を紹介する文書。

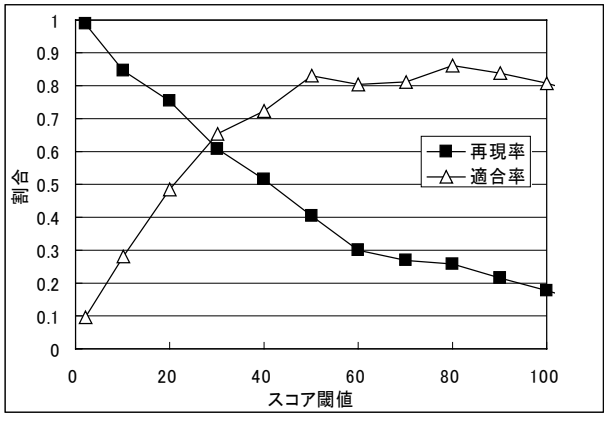

図 6 :クレーム抽出結果

7.3. トレンド分析部

クレーム急増検知の有効性を検証するために、特定の製品に関する風評情報に関して、実際の Web 文書を収集してクレーム抽出を実施し、クレーム情報のマクロな変動を調査した。

図 7は、発売後に不具合が発覚したある製品 X に関するクレーム急増の実際の分析例である。図中、横軸は調査期間を表わし、縦軸は一定期間内に抽出されたクレーム文書のスコア合計値を表わしている。

本事件では、メーカー側の不具合対応が悪いために Web 上でクレーム情報が急速に広まり、新聞で報道されるに至っており、図 7では、製品 X が発売されて不具合が発覚し、以降にクレームが急増していることが分かる。本試作システムのトレンド分析機能を用いることにより、新聞報道日以前にクレーム情報の広がりを検知して、危機の予兆を事前に把握することができるようになると考えられる。

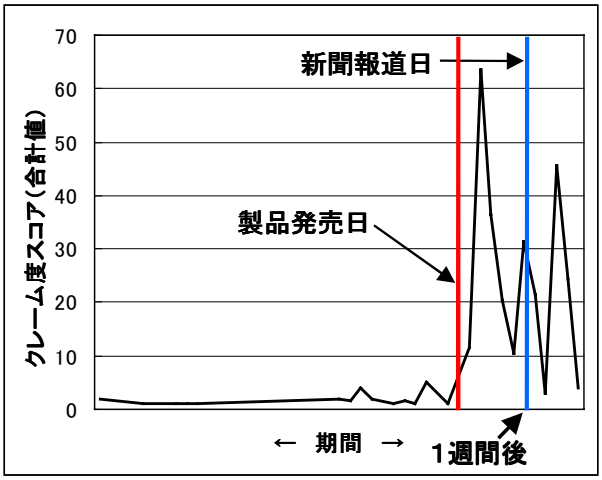

図 7:クレーム急増検知の例

8. おわりに

クレーム抽出技術を Web 文書に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] トラフィック・ワン・コミュニケーションズ(TOCC)社 ホームページ：
http://www.tocc.co.jp/search/

[5] 諸橋, 他 “テキストマイニング：膨大な文書データからの知識獲得 ― 意図の認識 ―, ” 情報処理学会 第 57 回(平成 10 年後期) 全国大会 3-75, 1998.

[6] “日本語完全対応 e メール自動分類・配信ソリューション ― MatchMail-CallCenter ―,” ビジネスコミュニケーション Vol. 37, No. 5, 2000.

[7] 立石, 他 “インターネットからの評判情報検索,” 情報処理学会 研究会資料 (NL 144-11), pp. 75, 2001.

[8] ガーラ社 ホームページ：
http://www.gala.jp/

[9] デジタルアーツ社 ホームページ：
http://www.daj.co.jp/