2E1-02

# 感情表現と用語のクラスタリングを用いた時系列テキスト集合からの話題検出

A method for detecting topics based on sentiment expressions and word clustering

福原 知宏[*1]
Tomohiro Fukuhara

中川 裕志[*2]
Hiroshi Nakagawa

西田 豊明[*3]
Toyoaki Nishida

[*1] 科学技術振興機構 社会技術研究開発センター
RISTEX, JST

[*2] 東京大学情報基盤センター 図書館電子化部門
University of Tokyo

[*3] 京都大学大学院 情報学研究科
Kyoto University

A method for detecting social events based on sentiment expressions and word clustering is proposed. Proposed method detects social events associated with specific sentiment or topic. The method also detects topics by classifying words into clusters based on correlation of occurrence over timeline, and co-occurrence of words in a text. Overview of the method and some analysis results obtained from proposed method are described.

## 1. はじめに

今日, 計算機の普及に伴い膨大な量のテキストが日々流通され蓄積されるようになった. こうしたテキストを分析することで, 社会においてどのような出来事が問題になっているかを把握できる. 人々がどのような出来事に対してどのような感情を抱いたかを知ることは企業や行政のサービス向上にとって重要である. 一方, 社会の出来事と市民感情の関係は過去の新聞記事を読み返せば分かるが, 記事の量が膨大であり, 人手での作業には不向きである.

本論文では感情表現と用語のクラスタリングを用いた時系列テキスト集合からの話題検出手法を 3 点提案する. 提案手法は日付情報を含むテキスト集合を入力とし, (1)ある感情について話題の推移を示すグラフ (話題グラフ) を生成する手法, (2)ある話題について感情の推移を示すグラフ (感情グラフ) を生成する手法, (3)キーワードの時間軸上の相関関係とテキスト内の共起情報を用いてキーワードをクラスタリングし話題ごとにキーワード

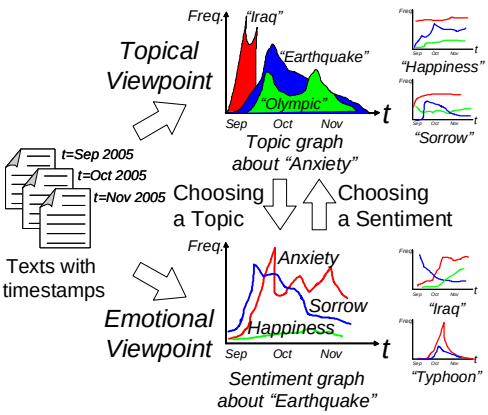


**Figure 1. 提案手法の概要**
**(感情表現を用いた話題検出手法)**

連絡先:東京大学人工物工学研究センター,
柏市柏の葉 5-1-5, phone: 04-7136-4275,
e-mail: fukuhara@race.u-tokyo.ac.jp

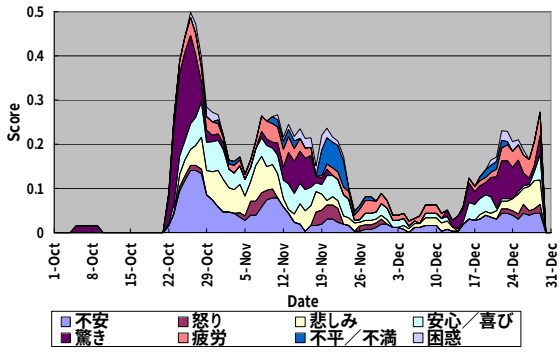


**Figure 2. 感情グラフの例 (2004/10/1 から 2004/12/31 までの間で " 地震 " と共起した感情カテゴリ)**

を可視化する手法の 3 点である.

Figure 1 に(1), (2)に関する提案手法の概要を示す. Figure 2 に話題グラフの例を, Figure 3 に感情グラフの例をそれぞれ示す. Y 軸は感情表現あるいは共起語の出現頻度を示す得点である. 両グラフとも朝日新聞 2004 年版を対象に実験を行った結果である. Figure 2 は新潟中越地震の発生した 2004 年 10 月 1 日から 12 月 31 日までの間で"地震"と共起した感情表現のカテゴリである. 時間経過とともに共起する感情カテゴリが変化していることがわかる. また Figure 3 はアテネオリンピックの行われた 2004 年 8 月 1 日から 8 月 31 日までの間で, "安心／喜び" の感情表現と共起した用語の推移を示している. "安心／喜び" の表現はこの時期, オリンピック関連の用語と結びついて用いられていたことが分かる.

本論文の構成は次のとおりである. 2.では提案手法の概要について述べる. 3.では提案手法の実行例について述べる. 4.では本論文のまとめと今後の課題について述べる.

## 2. 提案手法の概要

本節では提案手法の概要として, (1)感情表現の収集, (2)キーワード抽出手法, (3)感情表現を用いた話題検出手法, (4) 2.4 用語のクラスタリングによる話題検出手法について述べる.

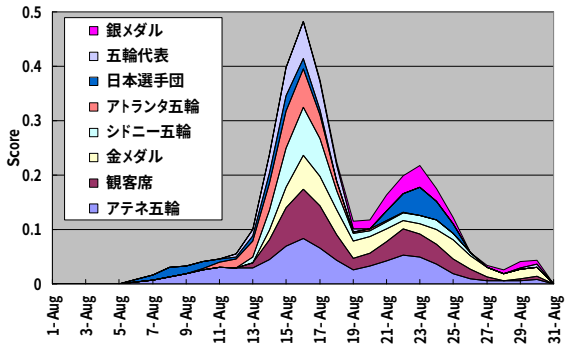


**Figure 3.“安心／喜び”の感情表現と共起した用語（2004/8/1 から 2004/8/31 まで）**

## 2.1 感情表現の収集

提案手法は感情表現に基づいてテキスト集合を解析する．筆者らは提案手法で用いる感情表現を手動で新聞記事より収集した．感情表現の収集には 2003 年と 2004 年に発生した化学事故に関する新聞記事から 8 カテゴリ 383 個の感情表現を抜き出した．

Table 1 に抽出した感情表現の例を示す．対象となった新聞紙は全国紙 4 紙（朝日新聞，読売新聞，毎日新聞，日本経済新聞），地方紙 3 紙（北海道新聞，下野新聞，中日新聞）である．

**Table 1. 感情カテゴリと感情表現**

| 感情カテゴリ | 感情表現 |
|---|---|
| 不安 | ‘心配’, ‘気がかり’, ‘怯えて’, ‘怖い’, ‘一喜一憂’, ‘気に掛け’, ‘複雑な心境’, ‘懸念’, ‘憂慮’, ‘心痛’ |
| 悲しみ | ‘悲しい’, ‘涙’, ‘目頭を押さえ’, ‘すすり泣く’, ‘嗚咽’, ‘嘆く’, ‘悲嘆’, ‘意気消沈’, ‘悲痛’, ‘号泣’ |
| 怒り | ‘怒る’, ‘怒り’, ‘憤る’, ‘憤り’, ‘苛立つ’, ‘許せない’, ‘厳しく非難’, ‘語気を強めた’, ‘声を荒げた’ |
| 安心／喜び | ‘安心’, ‘嬉しい’, ‘ほっとした’, ‘安堵した’, ‘声を弾ませた’, ‘笑顔’, ‘明るい表情’, ‘笑った’ |
| 困惑 | ‘苦悩’, ‘苦渋’, ‘困惑’, ‘当惑’, ‘戸惑い’, ‘頭を抱えた’, ‘頭を悩ませ’, ‘落胆’, ‘放心’, ‘呆然’, ‘ため息’, ‘やりきれない’ |
| 疲労 | ‘疲れた’, ‘疲労’, ‘疲れ果てた’, ‘ぐったり’, ‘落胆’, ‘がっかり’, ‘放心’, ‘意気消沈’, ‘あきらめた’, ‘挫折’, ‘途方に暮れて’, ‘遺憾’, ‘無念’ |
| 不平／不満 | ‘不平’, ‘不満’, ‘不服’, ‘不公平’, ‘納得できない’, ‘ぶ然’, ‘眉をひそめた’, ‘苦り切る’, ‘砂を噛む’, ‘難色’, ‘渋い表情’ |
| 驚き | ‘衝撃’, ‘驚いた’, ‘取り乱した’, ‘混乱’, ‘動揺’, ‘狼狽’, ‘あぜん’, ‘慌てた’, ‘動転’ |

## 2.2 キーワード抽出手法

話題を示すキーワードの抽出が重要である．我々は中川らの提案している重要語抽出システム：言選 Web[Nakagawa2003]を用いた．言選 Web は通常の形態素解析器とは異なり複数の単語からなるキーワードを単語間のつながりを計算することで抽出する．提案手法では重要語の特徴量に言選 Web の得点を用いる．

## 2.3 感情表現を用いた話題検出手法

感情表現を用いた話題検出手法（話題グラフと感情グラフ作成手順）の概要を次に示す．

### (1) 話題グラフ作成手順

利用者は事前に感情カテゴリ $s$ と検索期間 $D=(d_1, d_2, \ldots, d_l)$ を指定するものとする．

1. 検索期間 $D$ 中のそれぞれの日 $d_i(i=1,2,\ldots,l)$について，感情カテゴリ $s$ に含まれる感情表現を含む記事を検索する．
2. 検索された記事に言選 Web を適用し，キーワード集合を得る．
3. 各キーワード $w_j(j=1,2,\ldots,N)$について，感情表現集合との共起度の平均値 $c$ を求め，この $c$ を日付 $d$ における感情カテゴリ $s$ と $w_j$ の共起度とする．なお共起度は記事内共起として計算し，共起度には Dice 係数を使用した．
4. 各キーワードの期間中の重要度を計算し，重要度上位 n 語を取り出す．重要度計算には(1)キーワードの出現日数，(2)出現日数についての逆文献頻度，(3)言選 Web の得点の積を用いた．
5. 取り出した n 個のキーワードについて，共起度の時間的推移をグラフとして出力する．なお見やすさのため前後 1 日計 3 日間の移動平均を適用する．

### (2) 感情グラフ作成手順

利用者は事前にキーワード $w$ と検索期間 $D=(d_1, d_2, \ldots, d_l)$ を指定するものとする．

1. 検索期間 $D$ 中のそれぞれの日 $d_i(i=1,2,\ldots,l)$について，キーワード $w$ を含む記事を検索する．
2. 検索結果に含まれる記事集合を対象として，各感情カテゴリについて感情表現の出現頻度の和を取る．
3. 出現頻度の時間的変化をグラフ出力する（見やすさのため，前後 1 日計 3 日間の移動平均を適用する）．

## 2.4 用語のクラスタリングによる話題検出手法

提案手法は(1)用語出現頻度の時間軸上の相関と(2)テキスト内での共起関係を見ることで用語をクラスタリングする．手法の概要は次のとおりである．なお利用者は事前に検索期間 $D=(d_1, d_2, \ldots, d_l)$ を指定するものとする．

1. 検索期間 $D$ に出現する記事から言選 Web を用いてキーワードを抽出する．
2. 各キーワードについて言選キーワードのスコアの総和順に上位 n 語を抜き出す．
3. 選出した n 語を対象にキーワード同士の時間軸上の相関を求め，相関係数が閾値 $t_1$ 以上の組を抽出する
4. 抽出した用語について互いに記事内共起を求め，閾値 $t_2$ 以上の組を抽出する
5. 抽出した用語の組について時間軸上と 2 次元平面上に展開する

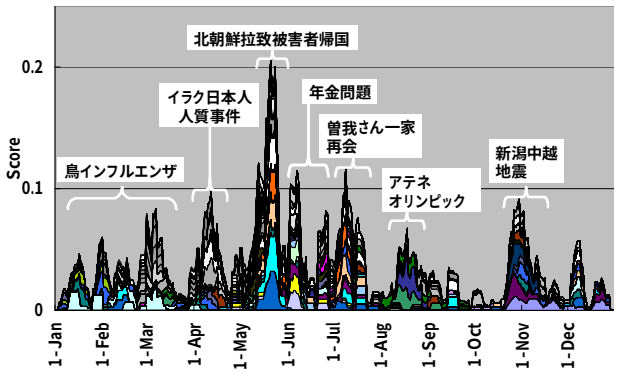


**Figure 5. 話題グラフの例**
**（“不安”カテゴリとの共起語, 2004 年）**

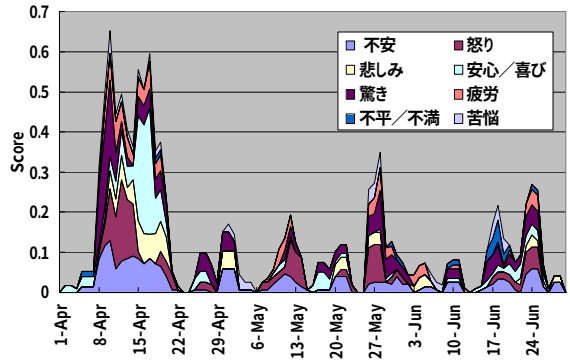


**Figure 6. 感情グラフの例（“イラク”と共起する感情カテゴリの推移（2004/4/1 から 2004/6/30 まで））**

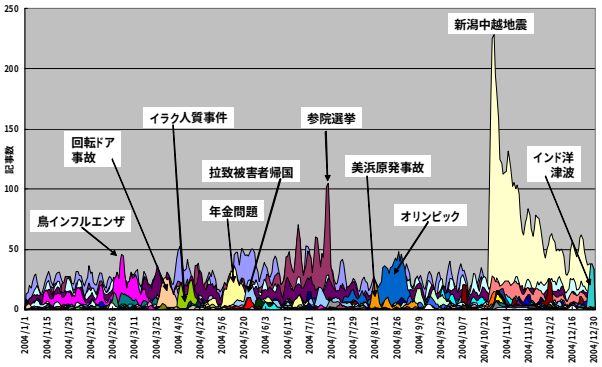


**Figure 7. クラスタリング結果 I：時間軸上での話題の分布**

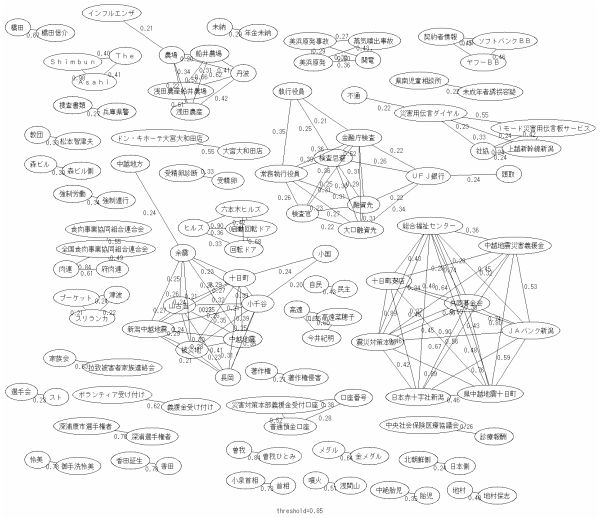

**Figure 8. クラスタリング結果 II: キーワードのネットワーク**

## 3. 実行例

本節では提案手法の実行例について述べる.

### 3.1 感情表現を用いた話題検出

#### (1) 話題グラフ

Figure 5 に検出した話題グラフを示す. グラフは 2004 年1年間における“不安”カテゴリの感情表現と共起したキーワードの推移を示している. キーワードを手動で話題ごとに分類すると, 鳥インフルエンザ(1 月から 3 月), イラク日本人人質事件(4 月), 北朝鮮拉致被害者帰国(5 月), 年金問題(6 月), 曽我ひとみさん一家再開(7 月), アテネオリンピック(8 月), 新潟中越地震(11 月)等の分類が可能であった. このように話題グラフでは, 感情と共起関係にあるキーワードの推移を確認できる.

#### (2) 感情グラフ

感情グラフの例として Figure 6 に 2004 年 4 月から 6 月までの間において, “イラク”と共起した感情カテゴリのグラフを示す. 事件の進展に伴って“イラク”と共起した感情表現は変化していたことが分かる. 特に 4 月前半は日本人が拘束されたことに間毛する“驚き”カテゴリと, 人質解放に伴う“安心／喜び”カテゴリが目に付く. このように感情グラフによって, ある話題がどのように社会に映ったかを把握できる.

### 3.2 用語のクラスタリングによる話題検出

Figure 7 と Figure 8 にクラスタリングによる話題検出の結果を示す. Figure 7 は時間軸上での話題の分布である. Figure 5 と比べると, クラスタリングの結果, 個々のキーワード単位ではなく, より大きなまとまりである話題を単位としてクラスタを見ることができる. Figure 8 はクラスタを 2 次元平面上に投影したグラフである. 同じ話題のキーワードがクラスタを形成している様子が分かる. このように時間軸上の相関関係と記事内での共起関係により, キーワード集合を話題単位にまとめることができる.

## 4. まとめ

本論文では感情表現と用語のクラスタリングを用いた時系列テキスト集合からの話題検出手法を 3 点提案した. 今後の課題として, (1)得られた結果の評価, (2)感情表現の自動獲得, (3)新聞記事以外のテキストへの適用について取り組む.

## 参考文献

[Nakagawa2003] H.Nakagawa and T.Mori.: Automatic Term Recognition based on Statistics of Compound Nouns and their Components. Terminology, Vol.9 No.2, 201-209, 2003.