# キーフレーズ抽出と帰納学習を用いたテキスト分類

1Q−4

酢山 明弘　櫻井 茂明　市村 由美　折原 良平

(株)東芝 研究開発センター

## 1. はじめに

オンラインでのアンケートの増加に伴い，比較的多量のアンケート回答を短時間で獲得することができることになった反面，高速にアンケートを分析・集計することが求められている．

回答欄が定型フォーマットであるならば，機械的に自動認識して集計することは容易である．しかし，自由記述回答欄の場合はアンケート分析者の手作業により分析・分類・集計がなされるのが一般的であり，分析のボトルネックとなっている．また，自由記述回答欄によるアンケートの分析で重要なタスクの1つとして，回答者の苦情や要求に対し，必要性の高いものから即座に対応する必要がある．

本論文では，自由記述回答に関する上記タスクを支援することを目的に，アンケート中の自由記述回答文から苦情文や要求文など任意のカテゴリに属するアンケートのみを抽出・提示する方法を提案する．とくに，自由記述回答文の分類に必要な特徴を抽出する方法と，帰納学習による分類規則発見法について述べる．

## 2. 自由記述回答文のフィルタリング

自由記述回答を特徴づける表現は，そのほとんどが文末に現れる．例えば「～ほしい。」ならば要求，「～すぎる。」「～が悪い。」「～がダメ。」なら苦情である．このような表現を辞書に登録して苦情文や要求文を抽出する方法もあるが，それら目的とする回答文以外の回答文を抽出しないための登録表現の記述方法や，目的とする回答文すべてを抽出するために必要な登録表現数が多いと考えられるため，負荷が高いといえる．

したがって，上記表現をあらかじめ辞書に登録することなく，自由記述回答文から分類を特徴づける表現を抽出して，自動的に分類規則を学習する過程を加えた自由記述回答文のフィルタリング法（図1）が望まれる．

図1のフィルタリング法においては，あらかじめシステム設計者により，分類を特徴づける表現を抽出するためのルールであるフレーズ抽出ルールが与えられている．もし，分類規則が1つも存在しない場合には，アンケート回答文データベース内の自由記述回答文の部分集合がシステムによってランダムに選択され，各回答文が区切り文字により文単位にセグメント化される．セグメント化された各文に対して，アンケート分析者がカテゴリを与え訓練回答文とする．

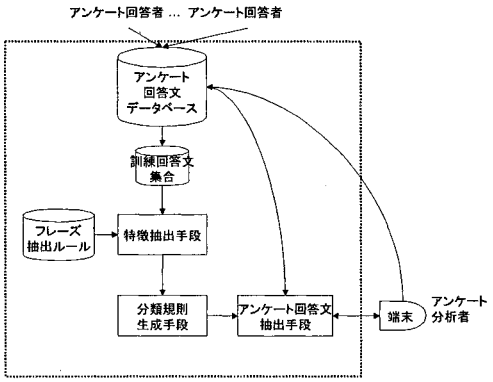


図1　自由記述回答文のフィルタリング

特徴抽出手段では，まず訓練回答文に対してシステム設計者により与えられたフレーズ抽出ルールを適用して，自由記述回答文を特徴づける表現集合（キーフレーズベクタ）を抽出する．次にキーフレーズベクタを用いて訓練回答文を帰納学習可能な表現（アンケート回答文特徴ベクタと呼ぶ）に変換する．

そして，分類規則生成手段において，アンケート回答文特徴ベクタを帰納学習することで分類規則を得る．

特定のカテゴリに属する自由記述回答文を抽出する場合は，自由記述回答文を分類規則により推論させカテゴリを決定する．ただし，自由記述回答の表現は多様であるため，初期分類規則だけでは一般的に適切に分類することはできない．よって，アンケート分析者が見て，明らかに分類結果が間違っていると判断できる自由記述回答文に関しては，インタフェースを通じて適切なカテゴリに修正することで，分類規則を再学習する．

以下では，本論文の特徴である特徴表現抽出と分類規則の学習に関して具体例を用いながら説明を行う．

### 2.1 自由記述回答文からの特徴表現抽出

表1　フレーズ抽出ルール　例

| *ID* | *ルール* |
|---|---|
| #R1 | {<名>\|<両名>\|<固固>\|<人固>\|<姓固>\|<地固>\|<名固>}{の\| }<英> |
| #R2 | {<五．>\|<下一>\|<上一>}{ \| \+<付>}\+<活尾> |
| … | … |

Text Categorization Using Key Phrase Extraction and Inductive Learning
Akihiro SUYAMA, Shigeaki SAKURAI,
Yumi ICHIMURA, Ryohei ORIHARA
Corporate Research & Development Center,
Toshiba Corporation.

表2 訓練回答文 例

| ID | AID | 自由記述回答文 | 分類カテゴリ |
|---|---|---|---|
| #T1 | #Q3#1 | 製品 O はよく壊れて困る。 | 苦情 |
| #T2 | #Q3#2 | 製品 O はデザインよい。 | 感想(良) |
| #T3 | #Q3#2 | けど、すぐ壊れて困る。 | 苦情 |
| #T4 | #Q4#1 | 製品 N は性能がよい。 | 感想(良) |
| #T5 | #Q4#1 | が、デザイン悪い。 | 感想(悪) |

表1はフレーズ抽出ルールを示す．フレーズ抽出ルールは ID と形態素の正規表現であらわされたルールから成る高々十数個の規則であり，語尾表現，文末動詞句表現および名詞句表現をキーフレーズとして抽出する．表1のフレーズ抽出ルール＃R2 は次のように解釈する．連続する2つまたは3つの形態素からなり，第1形態素として五段活用動詞または下一段活用動詞または上一段動詞のいずれかがあり，第2形態素として活尾あるいは第2，3形態素として付属語・活尾となる連続した形態素列からなる回答文上の表現を抽出する．

表2は，訓練回答文の例である．訓練回答文は訓練回答固有の ID，AID(質問番号とセグメント番号によりラベルづけ)，セグメント化された自由記述回答文，およびアンケート分析者が与えた分類カテゴリから成る．例えば，表2の＃T1``製品Oはよく壊れて困る"を形態素解析すると，"/製品<名>/O<英>+は<付>/よ<形>+く<活尾>/壊れ<下一>+て<付>/困<五ら>+る<活尾>/。<句読>"となり，表1のフレーズ抽出ルール＃R1，＃R2から，キーフレーズとして"/製品<名>/O<英>"および"/困<五ら>+る<活尾>"を得る．同様の処理を他の訓練回答文に関しても行い表3のキーフレーズベクタを得る．

表3 キーフレーズベクタ 例

| ID | キーフレーズ |
|---|---|
| #K1 | /製品<名>/O<英> |
| #K2 | /困<五ら>+る<活尾> |
| #K3 | /よ<形>+い<活尾> |
| #K4 | /デザイン<両名>/よ<形>+い<活尾> |
| #K5 | /製品<名>/N<英> |
| #K6 | /性能<名>+が<活尾>/よ<形>+い<活尾> |
| #K7 | /デザイン<両名>悪<形>+い<活尾> |
| #K8 | /悪<形>+い<活尾> |

自由記述回答文特徴ベクタをキーフレーズベクタの先頭から順にフレーズの出現／非出現により1／0を与えたベクトル表現とした．したがって，＃T1のアンケート特徴ベクタは"11000000"となる．

## 2.2 分類規則の学習

表4 アンケート回答文特徴ベクタ集合 例

| ID | AID | 特徴ベクタ | 分類カテゴリ |
|---|---|---|---|
| #F1 | #Q3#1 | 11000000 | 苦情 |
| #F2 | #Q3#2 | 10110000 | 感想(良) |
| #F3 | #Q3#2 | 01000000 | 苦情 |
| #F4 | #Q4#1 | 00101100 | 感想(良) |
| #F5 | #Q4#1 | 00000011 | 感想(悪) |

2.1 節により得られたアンケート回答文特徴ベクタ(表4)は，キーフレーズベクタ(バイナリ属性集合)によって表現された帰納学習の訓練事例といえる．アンケート回答文特徴ベクタ集合は，表2の訓練回答文集合の自由記述回答文フィールドをキーフレーズベクタの出現／非出現を列挙した特長ベクタフィールドに置き換えた構成から成る．

帰納学習法としては，高速かつ強力な学習手法であることから，ファジィ決定木を学習する IDF[櫻井 1996]を適用した．また，IDF は分類クラスにおけるあいまい性を(確信度つき分類クラスとして)明示的に表現する点でも，自由記述回答文のカテゴリ判断のようなあいまい性をもつ問題にも適しているといえる．

## 3. 実験

2節で述べた自由記述回答文フィルタリング法を実装し，自由記述回答文 1118 件に対し以下の評価実験を行った．まず，すべての回答文にカテゴリを与え，苦情文・要求文抽出の再現率・正解率の測定を 10-クロスバリデーション[Weiss 1991]で行った．カテゴリとして苦情(強)，苦情(弱)，感想(悪)，要求，問い合わせ，感想(良)，その他の7種類を用意した．

単語ベースで切り出した場合，再現率9%，正解率34%であったのに対し，キーフレーズベクタで抽出した特徴ベクタを用いた場合，再現率21%，正解率87%であり，本提案手法が有効であると判明した．しかしながら，本実験で用いたデータは数千件の回答文中，苦情や要求に該当するアンケート回答文は13件程度と非常に少ないため，アンケート分析者の立場からすれば再現率が高い方が有効であるといえる．

キーフレーズベクタ371属性には，苦情文・要求文のみを抽出するに十分なキーフレーズが含まれていることを確認している．すなわち，学習時において有効な属性を選択していないことが課題であり，キーフレーズ抽出ルールを洗練化することによる探索空間の絞りこみや，学習アルゴリズムの変更などを検討している．

分類タスクから見た場合における精度でも，単語ベースは平均 88%，キーフレーズでは平均 89.6%であることから，キーフレーズを特徴として抽出した方法の有効性を示すことができた．

## 4. おわりに

本論文では，自由記述回答文から苦情文や要求文など任意のカテゴリに属するアンケートのみを抽出・提示する方法を提案した．とくに，自由記述回答文の分類に必要な特徴を抽出する方法と，ファジィ決定木を用いた分類規則発見法について述べた．また，実験により提案手法の有効性を検証した．

## 参考文献

[Weiss 1991] Weiss, S, and Kulikowski, C., *Computer Systems that Learn*, San Francisco, Calif.: Morgan Kufmann, Publishers, Inc., 1991.

[櫻井 1996] 櫻井 茂明，荒木 大: ファジィ帰納学習アルゴリズムの改良，電学論 C, **116**, 9, 1057-1063, 1996.