

曖昧パターン照合方式を用いたテキストマイニング

相川 勇之 永井 明人 高山 泰博 今村 誠

三菱電機株式会社 情報技術総合研究所

1. はじめに

インターネットの普及に伴い、低年齢層、主婦層および高年齢者層まで含めた一般消費者による電子メールや掲示板の利用が増えている。これらのテキストには数多くの顧客の生の声が含まれており、商品開発やサービスの改善への活用が期待されている。そこで、量販店の販売情報などの大量データを分析する従来のデータマイニングに加え、大量のテキストから定性的なデータを抽出して分析するテキストマイニングが注目されている[1]。

我々は、同義性や類義性など単語や複合語間の潜在的な関係を自動的に抽出した概念索引を用いることを特徴とする概念抽出型テキストマイニング方式を開発した[2]。同方式は大量データの傾向分析に適しているが、単語共起に基づく統計手法のため、顧客満足度を示す「良い／悪い」などの表現を弁別しにくいという課題がある。また、営業日報における成功事例と失敗事例を分類するための手法として[3]が提案されているが、固定的なルール記述による抽出手法であるため、実用的な抽出精度を得るためには辞書登録を含めて大量のルール記述が必要という課題がある。

本稿では、これらの課題を解決するため、ルール記述の手間が少ない曖昧パターン照合方式を用いたテキストマイニング手法を提案する。製品アンケートを用いた実験の結果、提案方式により、好評・不評・要望などの情報を抽出する際の抽出精度向上に効果のあることがわかった。

2. 従来方式の課題

テキストマイニングの代表的な適用業務として、顧客アンケートの分析がある。アンケート分析において、顧客意見から顧客の満足度や要望を抽出して分析する機能が重要である。顧客アンケートの分析において重要な手順のひとつに肯定的意見(好評)と否定的意見(不評)の弁別がある。従来方式[2]は、単語共起に基づく

統計手法のため、顧客満足度を示す表現である「良い／悪い」のような反意語を弁別しにくいという課題がある。これは、共起単語が類似の出現傾向をもつため、「良い」と「悪い」の関連性も高いと判定されてしまうためである。

また、文献[3]で提案されている手法は、固定的なルール記述による抽出手法であるため、実用的な抽出精度を得るためには辞書登録を含めて大量のルール記述が必要という課題がある。

本稿では、これらの課題を解決し、好評・不評を表す表現をテキストから抽出して顧客アンケートの分析に活用するための曖昧パターン照合方式を提案する。同方式では、あらかじめ定義したパターン抽出ルールにもとづいて、分析のために必要な情報をテキストから抽出する。以下、3節で同方式によるテキストマイニング機能の概要について述べ、4節で同方式の効果を検証するための実験について述べる。

3. 曖昧パターン照合方式

この節では、顧客アンケート分析における顧客満足度分析(以下 CS 分析)機能を例に曖昧パターン照合方式の概要を説明する。

CS 分析機能とは、複数の単語に対して好評／不評などのスコアを定義し、これらの単語を顧客意見(自由記述)から自動抽出することによりスコアを付与し、図1に示すような分析結果を得られるようにする機能である。

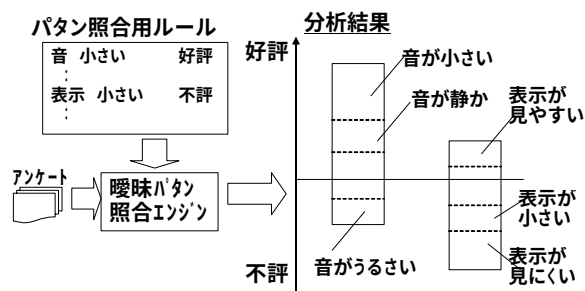


図1 CS分析機能の概略

図1には「音、うるさい」「表示、小さい」などの顧客の満足度に関する表現を抽出パターンとして定義しておき、これらの単語を自由記述中で検出して、これらに対応する好評／不評の数値を各アンケート回答に自動付与して顧客の満足度を分析する例を示している。

Text mining method using approximate pattern matching.
Takeyuki AIKAWA, Akito NAGAI, Yasuhiro TAKAYAMA,
Makoto IMAMURA
Information Technology R&D Center, Mitsubishi Electric
Corporation
5-1-1 Ofuna, Kamakura, Kanagawa, 247-8501, JAPAN.

パターン照合用ルールには、各パターンに対して、以下の照合オプションを指定できる。

- ・ 出現語順の無視などを指定できる語順指定（日本語の自由な語順にも容易に対応）
- ・ ワイルドカードなどによる表記ゆれ指定（送り仮名の有無などの表記ゆれに対応）
- ・ 否定、条件などを指定できる様相指定（条件表現と好評表現が連続する場合には、要望を意味することなどに対応）

これらの照合オプション指定により、少ないパターン記述で多くの表現に対応でき、情報抽出精度の向上が期待できる。

照合処理は、以下の手順で行う。まず、入力されたアンケートのテキストに対して形態素解析および文節解析を行ない、文節ごとに様相表現を抽出してデータを構造化する。次に、構造化されたデータに対してパターン照合用ルールを順次適用し、パターンに合致するデータから必要な情報を抽出する。

4. 実験

この節では、本方式の有効性を確認するために実施した製品アンケートに対する実験結果について述べる。

4.1. 実験方法

実験対象として、当社の製品に対するアンケートデータ 268 件を用いた。実験は以下の手順で行った。

- (1) 上記データをパターン抽出用の 160 件と、評価用 108 件に分割（ランダム抽出）
- (2) パターン抽出用のデータ 160 件から、肯定的意見／否定的意見／要望を抽出するためのパターン 122 件を手作業により作成
- (3) 評価用データ 108 件のうち、抽出すべき情報を含んでいる 99 件について正解データを作成（他 9 件については、製品の使用方法に関する単純な疑問点が記されているだけであつたため実験データから除外）
- (4) 上記パターンを用いて評価用データ 99 件から顧客満足度を抽出し、抽出結果を正否判定

4.2. 実験結果

表 1 に評価用データ 99 件に対する抽出実験の正解件数および不正解件数を示す。抽出精度は 83.8% だった。

表 1：実験結果

抽出項目	正解	不正解
肯定的意見/否定的意見	52	8
要望	31	8
合計	83	16

4.3. 考察

(1) 本方式の利点

本方式によれば、「～があれば便利」「選べたら絶対いい」など、条件を示す表現と好評を示す表現が連続して、これらをあわせて要望的な内容を示す表現が抽出できる。表 1 における「要望」の正解 31 件中、9 件が、このパターンにより抽出されており本方式による抽出精度向上の効果があつた。抽出表現の具体例としては、表 2 に示すような表現がある。

表 2：要望表現の例

「～があるとかかなり便利になると思う。」
「こんな機能があれば便利と思います。」
「もっと簡単に出せると嬉しい。」

辞書単語に「好評価」などのスコアを記述する従来の分析機能では、上記の表現については好評意見であると誤って判定されてしまう。本方式では、条件表現と好評を示す単語が文内で連続して出現しているという要望意見抽出用パターンの記述により、抽出精度を向上することができた。また、様相指定により多様な表現をすべて網羅せずにパターン照合用ルールを記述できるので、記述するパターンの数を抑制できた。

(2) 今後の課題

不正解 16 件のうち、文節解析における解析誤りがのべ 9 件あつた。このうち 6 件については、平仮名連続、口語体に対する解析誤りである。これらについては、口語対応を強化した解析用辞書の開発が必要である。

5. まとめ

ルール記述の手間が少ない曖昧パターン照合方式を用いたテキストマイニング手法を提案した。製品アンケートを用いた実験により、提案方式が抽出精度向上に効果のあることを示した。今後は、解析用辞書の口語対応強化を進めていく。

今回は実験対象としてアンケートデータを用いたが、パターン抽出用ルールを差し替えることにより、故障事例データの自動分類などにも本方式を適用できる可能性がある。これらの別データへの適用についてもあわせて検討していく予定である。

参考文献

- [1] 市村由美, 他: "テキストマイニング事例紹介", 人工知能学会誌, Vol.16, No.2. (2002)
- [2] 相川 勇之, 他: "概念抽出型テキストマイニングによるアンケート分析手法の提案", 情処 DD 研, DD38-1. (2003)
- [3] 市村由美, 他: "日報分析システムの開発", 信学技報 NLC2000-26. (2000)