アンケート自由記述のテキストマイニングに関する検討

4N－4

藤井洋一、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　音声・言語インタフェース技術部

1.はじめに

近年、アンケートに自由記述されたテキストやコールセンタにおけるヘルプデスク業務で蓄積された問い合わせ内容のテキストのように、蓄積されたテキスト情報を活用したいという要求が高まっている。しかし、従来のデータマイニングでは、テキストで自由記述された内容以外の数値化できるデータのみしか扱うことが出来なかった。

従来のテキストマイニング技術には、ヘルプデスク業務の問い合わせ内容に対して、テキスト中の重要語に関係する顧客の意図を抽出しようという試みがある[1]。この方法は、単語と概念を対応づけるカテゴリ辞書を用意して問い合わせ内容を解析し、用語の揺れを吸収し、テキスト中のモダリティや否定表現をパラメータ化して扱うことで問い合わせ内容のクラスタリングを行っている。さらに、テキスト中から抽出した単語、カテゴリがテキストを表現するとして、単語、カテゴリと、テキスト以外の定型情報との間でデータマイニングすることで問い合わせの特徴をつかもうとしている[2]。ヘルプデスク業務の場合には蓄積されるデータを日々分析するため、カテゴリ辞書を作成するコストはあまり問題にならない。

しかし、アンケート結果の分析は、蓄積されるデータを日々分析するのではなく、アンケートを集計した時点で高々数回の分析を行うのが一般的であり、あらかじめカテゴリ辞書のように詳細な情報を持った辞書を作成するには、コストがかかりすぎるという問題がある。また、分析したい内容もヘルプデスクの問い合わせ内容の分析ほど最初から明確にできないことが多く、カテゴリ辞書に用語とその意図をあらかじめ記述しておくことが困難である。

そこで、カテゴリ辞書を準備しなくても、アンケートに自由記述された内容をクラスタリングする方法として、自由記述中に出現した単語によるクラスタリングを検討した。

2.提案するテキストマイニング方式

本方式のテキストマイニングの特徴は、蓄積されたテキストの内容を単語集合で表現し、共通する単語集合をもつアンケート自由記述の内容をクラスタリングすることにある。ここでは、あらかじめ記述内容の予測が困難なデータを扱うため、実行結果を見た上で修正を加え、再度クラスタリングすることが容易である必要がある。単語の出現傾向によってクラスタリングを行い、用語辞書、ストップワード、およびパラメータを修正し、再クラスタリングできるものとする。

システム構成を図 1に示す。

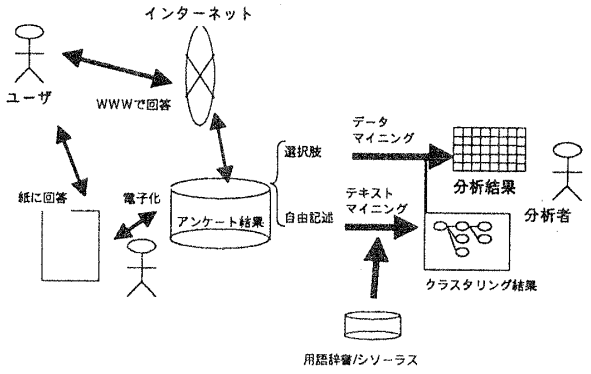

図 1　システム構成

インターネット上でのアンケートや、紙のアンケート結果を電子化したもので、テキストで回答される自由記述の部分が、テキストマイニングの対象である。本提案のテキストマイニングは、アンケートの自由記述部を対象として、自由記述中に出現した単語を利用してクラスタリングする。テキストマイニングした結果は、クラスタリングに使用した単語で特徴づける。この単語を利用することでアンケー

A Study of Text Mining for Free Text in Questionnaire
Youichi Fujii, Yasuhiro Takayama, Katsushi Suzuki
Human Media Technology Dept. Information Technology
R&D Center, Mitsubishi Electric Corp.
5-1-1 Ofuna, Kamakura, Kanagawa 247-8501, Japan

ト中の選択肢項目との間でデータマイニングを適用することも可能である。

2.1.本方式の処理の流れ

具体的なクラスタリング処理の流れを図 2に従って説明する。クラスタリングの処理は自由記述の設問毎に行う。あらかじめ用意する用語辞書／シソーラスは一般的なものを用い、必要に応じて追加するものとする。

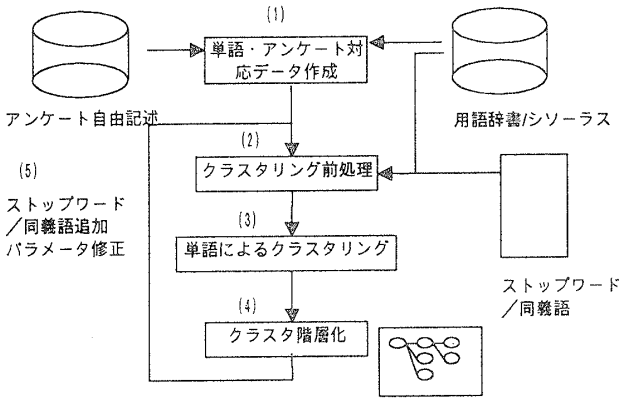

図 2　クラスタリングの流れ

(1)まず、対象となるアンケート結果の自由記述をを形態素解析し、単語をキーとして、単語が出現したアンケートをリストアップした単語・アンケート対応データを作成する。

(2) (1)で作成した単語・アンケート対応データに対して、ストップワードを処理対象から外し、同義語を１つの単語に同一視する処理を行い、単語を出現したアンケート数が多い順に並べる。

(3) (2)で並べた単語を高頻度の単語から順番にピックアップし、ピックアップした複数単語の相互情報量を計算し、相互情報量の値が一定の閾値以上の条件を満足する最大の単語集合を求める。求まった単語集合に対して、全ての単語を含むアンケートの集合をクラスタとする。この処理を繰り返す。

なお、相互情報量は以下の式で定義する。

$$\log_2\left(\frac{P(w_i \wedge \cdots \wedge w_j)/N}{(P(w_i)/N)\cdots(P(w_j)/N)}\right)$$

$P(w_i \wedge \cdots \wedge w_j)$は単語$w_i,\ldots,w_j$がすべて出現したアンケート数、Ｎはアンケート総数

パラメータとしては、クラスタの要素数、相互情報量の閾値、処理対象とする単語の最大数等を与える。

(4) 作成したクラスタに対して、クラスタリングに用いた単語集合の包含関係から階層を生成して表示する。クラスタのラベルとして、単語集合を表示する。

(5) 表示したクラスタリング結果の修正が必要なら、ストップワード／同義語の追加、パラメータの修正を行い、(3)からの処理を繰り返す。

2.2.本方式による効果

本方式のクラスタリングによって、アンケート結果中の自由記述部分に対して、単語集合によりアンケート結果の傾向をつかむことが可能となる。また、結果が好ましくなかった場合に、用語やパラメータを修正して、再実行することが容易である。

さらに、クラスタに対応づけられた単語を自由記述に対応づけることで、他の選択式のアンケート項目間、または、自由記述項目間で、データマイニングを行うことが可能である。

3.まとめ

今回、アンケート自由記述のテキストマイニングとして、自由記述中の出現単語によるクラスタリングを提案した。提案方式は、自由記述中に出現した単語について、相互情報量を利用してクラスタを生成し、単語集合の包含関係を元にクラスタを階層化する。本方式では、自動抽出した単語集合によりクラスタリングを行うので、カテゴリ辞書のような知識をあらかじめ用意しなくてよいのが特徴である。

今後は、本方式の有効性を試作し、評価する。

［参考文献］
[1]諸橋、那須川、長野「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回全国大会5K-3,第３分冊pp.75-76(1998).
[2]那須川、諸橋、長野「テキストマイニング：膨大な文書データからの知識獲得－概要－」情報処理学会第57回全国大会5K-4,第３分冊pp.77-78(1998).