青山学院大学　理工学部　情報テクノロジー学科

2007（平成 19）年度卒業論文要旨

意味解析を用いたテキストマイニングツール STM の適用性と分類精度の向上

片山　真紀子(15804024)　　永瀬　和彦(15804053)　　福田　雄介(15804061)

原　田　研　究　室

1.　はじめに

自由記述式のアンケートは選択式のアンケートに比べ、回答者の自由な意見を集約できるなどのメリットがある。しかし、大量のテキストデータを人手で分類し、分析するには多くの時間と人材の確保が必要である。そのため、近年注目を浴びているのがテキストマイニングである。既存のテキストマイニングでは形態素の表層的情報を中心とした解析が行われているが、結果として語意や語間の関係が把握できない、「何がどうだ」「何がどうした」などの複数の語からなる関係を把握できない、表現の揺れによる差異を吸収できないといった問題があった。

このような背景を踏まえ、昨年度原田研究室では、意味解析システム SAGE[1]を用いたテキストマイニングツール STM2006[3]を開発した。STM では日本文を意味グラフに展開し、対応する節同士の概念的な類似度や節間の深層格の類似度をベースに、類似部分グラフの大きさで 2 文の類似度を計測することで、表現が異なっていても同様な趣旨を持つ文を同意見として集約し分類する。本研究では STM の適用性と分類精度のさらなる向上を目指す。

2.　STM における処理概要

本システムでは図 1 に示すように、入力された CSV 形式のアンケートデータに対して意味解析システム SAGE を用い意味解析を行い、その結果を用いて句(個々の述語節とそれが伴う深層格を構成する節の集まり)を作成する。意味解析の結果や作成された句を SQL サーバに保存し、このデータベースを基に頻度分析、クラスタリング分析、時系列分析、コレスポンデンス分析を行う。クラスタリングの対象は句、文と文章である。

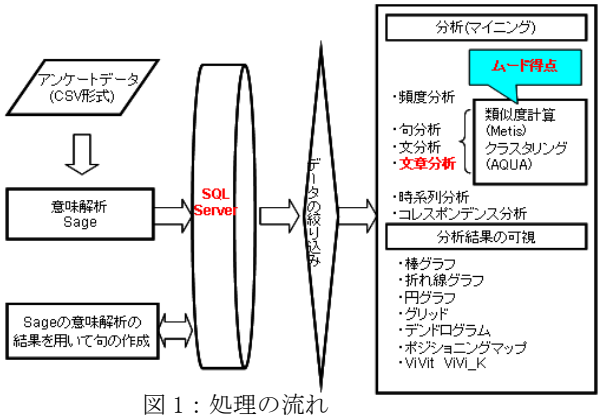

図 1：処理の流れ

本システムのユーザインタフェースの設計においては、1 分析に対して 1 画面を提供することで、複数の分析結果の同時比較を可能とした。

3.　人間の感性に近い文類似度の算出

3.1.　文節数に応じた文類似度の細分化

アンケートの意見分類では、ノードグラフ類似度単独よりも「何がどうだ」「何がどうした」といった深層格の類似度を含むアークグラフ類似度が重視される。

そこで本システムでは下式を用いて文類似度(グラフ類似度)を算出していたが、比較対象の意味グラフが両方とも 1 文節であるときや片方のみ 1 文節であるときは自動的にアークグラフ類似度が 0 になるため、他の文間の類似度がそれほど高くない場合にギャップが生じていた。そこで図 2 に示すように、文類似度を調整するための文節数に応じた細分化を行った。

$$\text{文類似度} = (1-\alpha)\cdot\text{ノードグラフ類似度} + \alpha\cdot\text{アークグラフ類似度}$$

α：関係重視率

(ⅰ) 両グラフとも1文節の場合
　文類似度＝(1－α / 5)・ノードグラフ類似度
(ⅱ)片方のグラフのみ1文節の場合
　文類似度＝(1－α / 3)・ノードグラフ類似度
(ⅲ)両グラフとも多文節の場合
　文類似度＝(1－α)・ノードグラフ類似度 + α・アークグラフ類似度

図 2: 文節数に応じた文類似度の細分化

3.2.　ムードによる調整

ムードとは、「事態や相手に対する話し手の判断や態度を表す文法形式」[2]であり、本システムでは SAGE による意味解析により該当語句へ付与される。ムードはアンケート回答における回答者の要求意図を表しているため、それらに得点を付与し文類似度へ反映させることで、より人間の感性に近い分類が可能となる。

ムードは 19 種類(なし、否定、禁止、要望、依頼、勧誘、願望、意志、困難、容易、可能、命令、当為、外観兆候、兆候、伝聞、推量、確信、過度)に分類される。今回の STM2007 では、否定、禁止、困難などの程度のグループと要望、勧誘、当為などの認識・態度のグループに分け、照合ノードペアに対して、グループ毎にお互いが持つムードの組み合わせで与えられる得点をムード類似係数として掛ける。1 つのノードが複数のムードを持つ場合は、平均のムード得点を掛ける。また、各照合ノードペアに対応するアークペアに関しては、両端のノードのムード類似係数を掛けることで調整を行う。

ノード類似度＝ノード概念類似度×*ムード類似係数*

$$\text{アーク類似度} = \text{アーク概念類似度} \times \frac{\text{両端のムード類似係数}}{2}$$

(例)説明が*分かりやすい*。[容易]
　内容が*分かりにくいようだ*。[困難、兆候]

図 3: ムードによる調整

4.　大容量データへの適用

従来のシステムには 2000 以上のデータを分析しようとすると、メモリが不足し分析できないという問題があった。そこで本システムではメモリを効率よく使用する方法に改善し、大容量のデータを分析することを可能にしている。さらに分析ごとに計算していた類似度をあらかじめデータベースに格納しておくことにより分析時間を短縮するように改善されている。

さらに STM で使用するデータベースを簡易的な Microsoft Access から SQLServer に変更した。高度な SQLServer を使用することにより、ADO.NET による高速なアクセスや、容量を 4GB まで増やすことを可能にしている。また Mirosoft Access では不可能であった 64bit のコンピューターでの使用も実現できる。

5.　文章分析

今回追加した機能の文章分析では、句・文分析の際に分割していた意見を分割せずにクラスタリングすることによって、個々の意見から新たな知識を獲得することができる。

本システムでは、アンケートデータの文類似度テーブルを用いて、新たに個々の意見の類似度の文章類似度テーブルをメモリ上に作る。それを用いてクラスタリングをする。

文章類似度を作る際、各々の意見を一文ずつ全文比較し、最大の組み合わせの平均を取ることで、個々の文章類似度として定義している。この類似度を用いてクラスタリングは文分析と同様の作業をしている。

図 4 : 文章分析のインターフェース

クラスタリングとして結果を表示する為、フォームの TreeView を用いなければならないが、文が長く表示しきれないので、詳細画面上に全文表示という項目で全文を表示させている。また類似度の最大値の組み合わせの平均を取っている為、クラスタリングの要因がわかりづらいので、最大値の組み合わせの中の最大の組み合わせを詳細画面上の一致文という項目でその組み合わせと類似度を表示させている。

6.　実験と評価

ムード値の内、要望、命令、依頼、当為といった希望的要因のムード値を 0.03 から 0.8 に上げた結果、図のような結果が得られた。今回は要望-当為、依頼-当為が大きく結果に表われている図である。

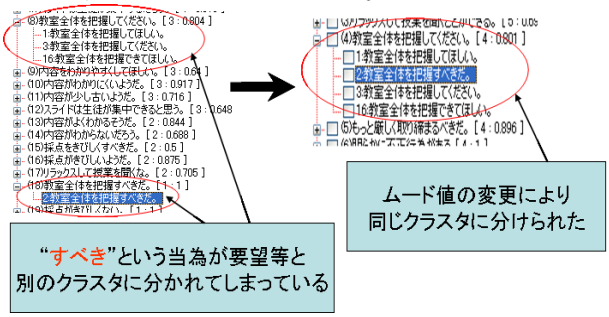

図 5 ：ムードによる改善結果

次に 4 章の大容量データへの適用による改善結果について述べる。システムの変更により大容量化を目指した結果、32 ビットのメモリ空間のコンピュータでは分析データの最大容量は 1000 近く増加し、3150 個のデータまで分析が可能となった。また 64 ビットのメモリ空間のコンピュータでは 7000 個のデータまで分析が可能となり、7000 人分のファーストフードに対するイメージのアンケートや、長い文章が多々含まれる web 上のクチコミの意見なども分析できるようになった。

さらにシステムを高速化した結果のデータ数による分析時間を以下のグラフに示す。

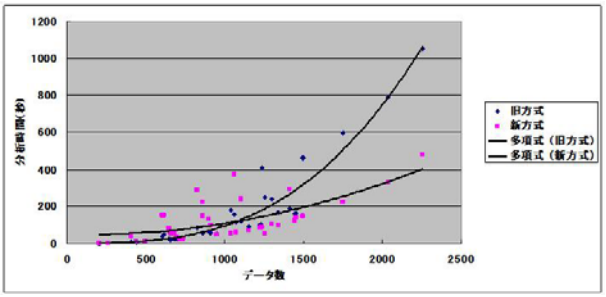

図 6 : データ数ごとの分析時間

図 6 の旧方式と新方式のグラフを見ると、データ数が増えるにつれ新方式の方が短い時間の分析で行えることが分かる。

【参考文献】

[1] 川口純一,青木洋,松田源立,原田実:"意味解析システム SAGE の精度向上"情報処理学会第 69 回全国大会論文集,1C-04,第 2 分冊 pp. 77-78. (2007.3).

[2] 佐藤直美,韓東力,原田実:”日本語意味解析に伴うヴォイス・テンス・アスペクト・ムードの決定”,情報処理学会第 67 回全国大会論文集,1J-03,第 2 分冊, pp.69-70 (2005.3).

[3] 西脇　剛,保立哲志,原田実:"意味解析に基づくテキストマイニングシステム STM"情報処理学会第 69 回全国大会論文集,2C-03,第 2 分冊 pp. 89-90. (2007.3).