2C-3

意味解析に基づくテキストマイニングツール STM

西脇　剛†　　保立　哲志†　　原田　実‡

†‡青山学院大学　理工学部　情報テクノロジー学科

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、本研究では原田研究室で開発した意味解析システム SAGE[1]を用いたテキストマイニングツール STM の開発を行う。STM では、日本文を意味グラフに展開し、2 つの意味グラフの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によって、表現が異なっていても同様な趣旨をもつ文を同意見として集約し分類する。

2.　システム概要

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 Access データベースに保存し、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文と句である。

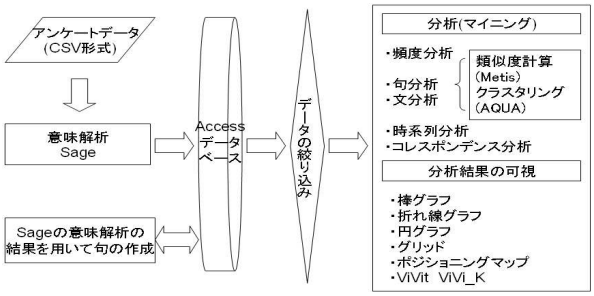

図 1　システムの構成

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。

Text Mining Tool STM based on semantic analysis

Tsuyoshi Nishiwaki†, Tetsushi Hotate†　and Minoru Harada‡

† Undergraduate School of Integrated Information Technology, Aoyama Gakuin University.

‡Department of Integrated Information Technology, Faculty of Science and Engineering, Aoyama Gakuin University.

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類似度が重視される。

そこで本システムでは(式 1.1)を用いて文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。そこで表 1 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。

文類似度＝(1－ α)・ノード類似度 + α・アーク類似度　(式1.1)
α：関係重視率

表 1　文節数に応じた細分化

(i) 両グラフとも1文節の場合
　文類似度＝(1－α / 5)・ノード類似度
(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)・ノード類似度 + α・アーク類似度

3.2. ムード得点による調整

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 8 つのグループ(否定系列、外要望系列、内願望系列、hardness 系列、easiness 系列、命令系列、推測系列、確信系列)に分類され、同一グループの照合ノードペアに対しては一致係数を掛け加点し、異なるグループの照合ノードペアに対しては不一致係数を掛け減点する。またムード得点グループが一致せず、(命令系列－外要望系列)、(hardness 系列－easiness 系列)、(推測系列－確信系列)の組み合わせとなった場合、不一致係数に加えて共鳴・反発係数を掛けることでノード類似度を調整する。また、各照合ノードペアに対応するアークペアに関しても、同様に調整を行う。

表 2　ムード得点による調整

ノード類似度=ノード類似度×一致係数(不一致係数)　(式2.1)
(例) 分かりやすく説明してほしい。　[外要望系列 / 要望]
　もう少し丁寧に説明してください。[外要望系列 / 依頼]
ノード類似度=ノード類似度×一致係数(不一致係数)×共鳴・反発係数　(式2.2)
(例) 説明が分かりやすい。[easiness系列 / 容易]
　内容が分かりにくい。[hardness系列 / 困難]

4. デンドログラムの有効活用

本システムでは、凝集型の階層クラスタリングを用いている。この手法は、N個のデータに対して1個のデータだけを含むN個のクラスタがある状態から、クラスタ間の距離関数に基づき、最も距離の近いクラスタを逐次併合する。デンドログラムとは、このクラスタリングの過程を示した樹形図のことである。本システムでは図2に示すように、全要素、クラスタ単位、各クラスタの要素単位の3つのデンドログラムの表示が可能である。これによりクラスタリング過程の全体像だけではなく、個々のデータの類似性を視覚的に捉え、細かな部分の関連性もつかむことができる。

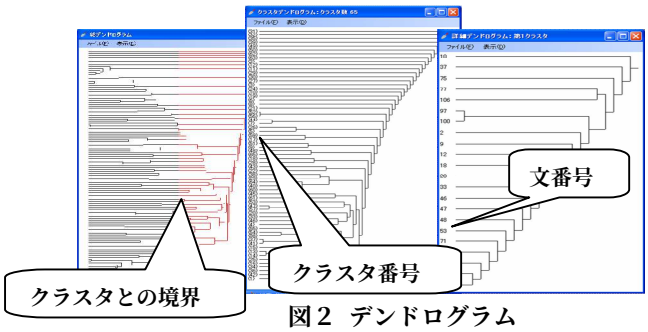

図2 デンドログラム

5. コレスポンデンス分析

コレスポンデンス分析では、クラスタリング分析の結果を用いて、アンケート回答者の属性と意見との相関関係をポジショニングマップに表現する。これにより、頻度分析などからは得られなかったデータ間の関連性などの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(文 or 句)のクラスタリング結果から、クラスタの要約文と要素数を回答者の属性ごとに分類したものをクロス集計表としてまとめ、それを基にポジショニングマップを作成する。回答者の属性は属性リストの項目を選択することで指定され、この項目に合ったデータ数がカウントされる。その後、作成されたクロス集計表のマトリックスよりポジショニングマップにおける回答者の属性および意見(クラスタの要約文)のプロット座標を算出し図3に示すように表示する。

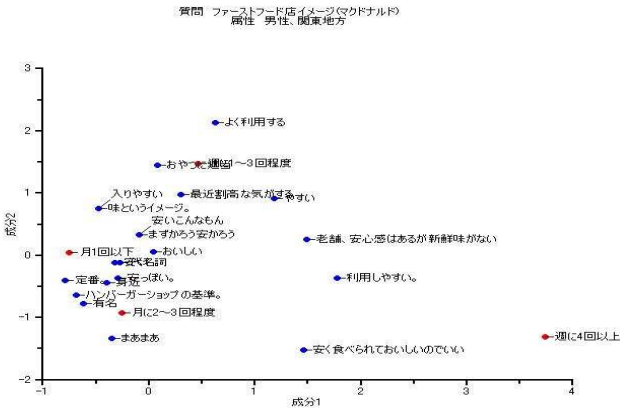

図3 ポジショニングマップ

ポジショニングマップでは、プロットの距離が近いほどデータ間の関連性が強く、遠いほど弱くなる。したがって、プロットの位置や距離から属性による意見の相違や、属性と意見との関連性を視覚的に捉えることができる。

6. 実験と評価

本研究では、意味解析の結果を用いた自由記述アンケートの分類を行った。図4は、「人工知能論の授業はどうでしたか」という質問文に対する回答81文を分類した結果の一部である。本システムの分類により、回答文群は52のクラスタに分類された(各クラスタは点線で区切られている)。

```
文30：スライドの授業は分かりやすかったです。
文54：スライドを用いて分かりやすかった。
文61：スライドを使って分かりやすかった。
文67：スライドが見やすくて分かりやすかった。
----------
文2 ：スピードが速い。
文65：ちょっとペースが速すぎて、途中からついていけなくなってしまった。
文68：進みが少し速かったので、当てられたときに考える時間がほしかった。
----------
文3 ：内容もわかりやすかった。
文4 ：ソフトウェア設計より内容が分かりやすかった
文64：内容に興味があったし、わかりやすかった。
----------
                :
----------
文49：もう少しテンポを落としてほしかったです。
文56：もう少し減らしてほしい。
文80：もう少しゆっくりやってほしかったです。
----------
                :
----------
文15：声が通って聞き取りやすかった
文46：マイクを使い声も大きく聞きやすかった。
----------
                :
----------
文25：課題が適度に出ていてよかったと思う。
文45：宿題、小テストの回数が妥当であったと思います。
----------
                :
```

図4 実験による分類結果(一部)

実験結果が示すように、同一クラスタ内の文は互いによく類似しており、クラスタ間には内容的な差があった。また、TF・IDF法を用いた商用のテキストマイニングツールによる分類と比較して、語意や語間の意味的関係を考慮した分類が可能となり、「もう少しテンポを落としてほしかったです」「もう少しゆっくりやってほしかったです」といった文が同一クラスタに分類されるなど、表現のゆれによる差異を吸収することができた。さらに、意味解析によって付与されたムードによって、「〜してほしい」「〜してもらいたい」といったアンケート回答者の要求意図を文類似度計算に反映させたことで、より人間の感性に合った分類結果を得ることができた。

【参考文献】

[1]原田実, 田淵和幸, 大野博之, "日本語意味解析システム SAGE の高速化・高精度化とコーパスによる精度評価", 情報処理学会論文誌, Vol.43, No.9, pp.2894-2902, (2002.9) .

[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", 青山学院大学大学院理工学研究科修士論文, (2004) .

[3]山本哲哉, 小林寛之, 米澤太一, "意味解析システム SAGE の精度向上と利便性の向上", 青山学院大学理工学部卒業論文, (2005) .