テキストマイニングのための
効果的なデータ収集方法に関する提案

代表者　鈴木裕利　　研究協力者　石井成郎

１．はじめに

これまで，先行研究において，「医療の現場で発生する情報に対するデータマイニング技術の応用に関する検討」を進めてきた[1]．対象とするデータには，身長，体温等の数値で表現される量的データと，看護記録等のテキストデータに代表される数値化できない質的データが存在する．量的データの分析については検定方法も確立されており，データマイニング技術の活用は進んでいる．一方，質的データは，量的データの分析方法では困難な心情・感覚等の解析が可能であると期待されて重要であると考えられているが，前処理の時間コスト，アンケートによるデータ収集には質問内容の工夫が必要な点等から，テキストマイニング技術を用いた質的データの分析の活用成功事例は少ないといわれている[2] [3]．そこで，先行研究での実験結果からアンケートを使用して質的データを収集する際に，キーワード系の回答を収集するには「リストアップ型」，心情・イメージ系の回答を収集目的とする場合には「品詞指定型」を指定することが望ましいとの知見が得られた．これは医療分野のデータに限定されない汎用的な知見といえる．

本プロジェクトでは，先行研究における知見を医療以外の場面に適用した実験を実施して，その有効性を検討する．また，その検討結果に基づき，より効果的な質的データの分析を実現する，テキストマイニングのための効果的なデータ収集方法に関するフレームワークの提案を目的とする．具体的には，汎用テンプレート，マニュアル等の作成により，テキストマイニング初心者にとって容易な分析を可能とする環境の構築を目指している．

本文では，前述の知見を検証するために実施した実験について報告する．

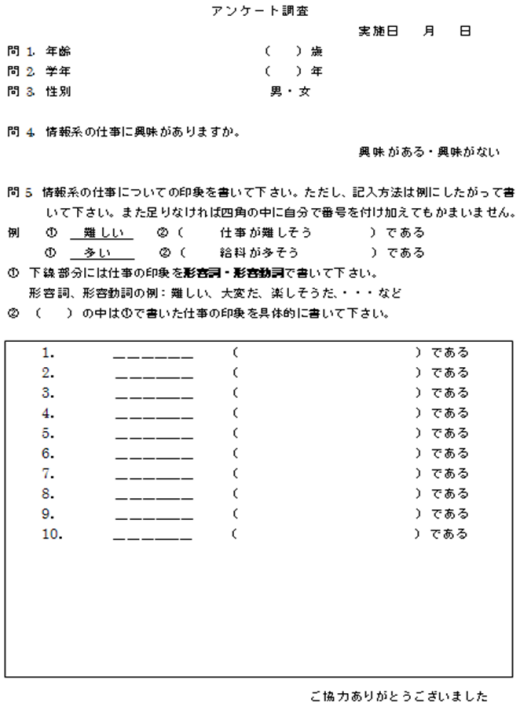
２．予備実験

前述したように，先行研究では「被験者がある質問に対してどのように思うのか，どのように感じるのか」等の心情・イメージ系の回答を収集目的とする場合には「品詞指定型」を指定することが望ましいとの知見が得られている．これは，有効な回答語数をより多く得られるという観点に基づいているが，さらに先行研究では，「品詞指定型」だけでなく「連言型」のアンケート用紙と組み合わせることでより効果的なデータの収集が可能ではないかと考察している．本予備実験では，先行研究で使われた「品詞指定型」と「連言型」の回答データを詳細に分析することにより，先行研究によって提案された回答形式の組み合わせの妥当性について分析する．ここでは，分析の一部の例を紹介する．

表１　出現形容詞と対応する連言型の文章

形容詞	対応している文章（抜粋）
難しい	ブラック（仕事が難しくてやばい） 難しそう 難しそう
忙しい	忙しそう 忙しい仕事
多い	多そう 資格が多い分野 残業の多い仕事 マイナスのイメージが多い デスマーチが多い仕事
深い（幅広い）	幅広い知識を持っている人が必要とされる
大きい（広い）	常に広く使われるもの ジャンルが広い職業。
しんどい、つらい（きつい	きつそう
長い	長く続かない。

先行研究ではマイニングした結果から，「多い」という形容詞が多く出現している．本アンケートでは，情報系の仕事に対するイメージがポジティブか，ネガティブかに分類して分析を行っているが，この「多い」という単語がポジティブかネガティブかについては，形容詞だけでは明確にならない．一方，連言型に記述された内容からは「資格が多い分野」，「残業の多い仕事」，「マイナスのイメージが多い」という回答となり，より詳細な内容が確認できる（表１参照）．よって，単語の意味からポジティブと判断した回答が，実際に連言型に書かれた内容からはポジティブな意味とは取れない場合があることが確認される．このように，得られる情報量については，品詞指定型に比べて連言型が多くなるといえる．本予備実験では，先行研究で収集した回答データに関して，得られる情報の内容について，品詞指定型と連言型との相違点を単語ごとに詳細な比較を行った．結果の詳細は省略するが，回答形式の組み合わせにより，より質の高いデータをより多く収集することが可能であることが確認された．


アンケート調査

実施日　月　日

問 1. 年齢　（　）歳

問 2. 学年　（　）年

問 3. 性別　男・女

問 4. 情報系の仕事に興味がありますか。

興味がある・興味がない

問 5. 情報系の仕事についての印象を書いて下さい。ただし、記入方法は例にしたがって書いて下さい。また足りなければ四角の中に自分で番号を付け加えてもかまいません。

例　① 難しい　②（　仕事が難しそう　）である

① 多い　②（　給料が多そう　）である

① 下線部分には仕事の印象を**形容詞・形容動詞**で書いて下さい。

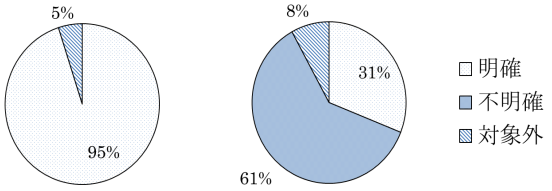
形容詞、形容動詞の例：難しい、大変だ、楽しそうだ、・・・など

②（　）の中は①で書いた仕事の印象を具体的に書いて下さい。

1. ＿＿＿＿＿（　）である
2. ＿＿＿＿＿（　）である
3. ＿＿＿＿＿（　）である
4. ＿＿＿＿＿（　）である
5. ＿＿＿＿＿（　）である
6. ＿＿＿＿＿（　）である
7. ＿＿＿＿＿（　）である
8. ＿＿＿＿＿（　）である
9. ＿＿＿＿＿（　）である
10. ＿＿＿＿＿（　）である

ご協力ありがとうございました

図 1　品詞指定改良型のアンケート用紙



品詞指定改良型　　品詞指定型

図 2　形容詞における回答データの質の比較

3．実験 1（品詞指定改良型の有効性の検証）

本実験では, 予備実験で確認された回答形式の組み合わせによるデータ収集の有効性の検証を行い, 評価を行うことを目的としている.

アンケート対象者は, 中部大学情報工学科の開講科目「企業情報システムと倫理」の受講生 63 名である．アンケートは,「情報系のイメージ」について質問している. 先行研究で使用した「品詞指定型」と本研究で提案する「品詞指定型」と「連言型」を組み合わせた「品詞指定改良型」との回答データの質を比較する（図 1,2 参照）.

図 2 より「品詞指定型」では回答が不明確な形容詞が多いが,「品詞指定改良型」では回答が明確な形容詞が多く出現していることが確認される. この結果から, 回答形式の組み合わせの妥当性の確認ができたといえる.

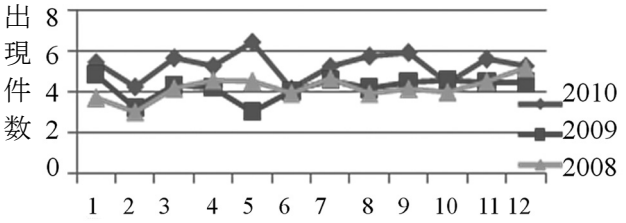


図 3　一人あたりの書き込み語種類数

4．実験 2（リストアップ改良型の有効性の検証）

本実験のアンケートの回答者は, 中部大学情報工学科の開講科目「創成 B」の受講者で，各年度約 100 名の受講者がある．授業への理解度，感想について，2008 年度は「自由回答型」, 2009 年度は紙媒体での指示による「リストアップ型」によって回答データが収集された．しかし，2009 年度の回答データの質にはバラツキが観察されたために収集方法に改善が必要とされた．よって，2010 年度は，新たにデータ収集のための web サイトを構築して「リストアップ改良型」として実施した. 本実験では, 改善後の方法について評価を行うことを目的としている.

結果からは, 一人当たりの書き込み語種類数において，2010 年度の「リストアップ改良型」は 2008 年度，2009 年度の結果に比べて多くなっていることが観察される（図 3 参照）．一方，総抽出語数，および，一人当たりの書き込み語数は 2009 年度の「リストアップ型」よりも減少していることが確認されている．この結果から,「リストアップ改良型」の回答形式の導入によって, 全体の書き込み語数に比較してマイニングに有効となる書き込み語種類数の割合が増加したことを意味しており, キーワード系の回答のデータの質が高くなったと考える.

3．おわりに

本文では, 回答形式の組み合わせによるデータ収集の妥当性の検証実験,「リストアップ型」のデータ収集方法の改善と検証実験について報告した. 今後は，本研究で得られた知見に基づいて，アンケート自動生成システムの設計, 実装を行う予定である.

参　考　文　献

[1]鈴木，石井,「テキストマイニング技術の効果的適用に関する提案」，情報科学リサーチジャーナル，Vol.17，pp.49-58，2010

[2] 那須川,「テキストマイニングの普及に向けて」-研究を実用化につなぐ課題への取組み-，人工知能学会誌，Vol.24，No.2，pp.275-282，2009.

[3] 大瀧，高橋，吉澤，今村,「テキストマイニングによる教育実習体験の分析」, 東京家政大学研究紀要, 第 50 集 1 号，pp.63-70，2010