

A-014

アンケート結果からの知識抽出に関する研究

Study About Knowledge Extraction From Questionnaire Result

岩切 広賢†
Hirotooshi Iwakiri

平山 正治†
Masaharu Hirayama

1. まえがき

本研究は記述回答と選択回答からなるアンケート結果に対して、要約処理と統計処理を行い、回答者全体の意見をまとめた結果を知識として自動的に抽出することを目的とした研究である。要約処理では、記述回答を入力し、「関心語とその評価」(知識1)を出力する。統計処理では、選択回答、関心語、および関心語利用状況を入力し、選択回答の各選択肢ごとに、特定の関心語に興味や関心をもつ回答者数を示す「関心度グラフ」(知識2)を出力する。以下では、上記の二つの処理からなる知識の自動抽出方法について述べる。

2. アンケート分析の課題と提案

従来のアンケート分析では、選択回答については選択肢ごとの統計をとり、グラフ化し、回答の分布を分析した[1]。記述回答については、人が自ら記述回答を読むことで、アンケート回答者の興味や関心を示す情報を得ていた(図1の上部参照)。このような分析によって、選択回答の分布と、回答者の興味や関心を知ることはできるが、二つの分析結果の相互の関連性からなる回答者全体の意見は容易に得られない。また、記述回答を人の手で分析することは非常に時間がかかり、分析者の能力や嗜好によって、分析結果が異なるという問題がある。

本研究は、名詞の重要度、記述回答の類似度、手がかり語の重要度などを用いた要約処理によって、アンケート回答者の「関心語とその評価」という知識を自動抽出する手法、および選択回答、関心語、関心語利用状況抽出を統計処理することによって、ある特定の関心語に対する「関心度グラフ」という知識を自動抽出する手法に関するものである。(図1の下部参照)

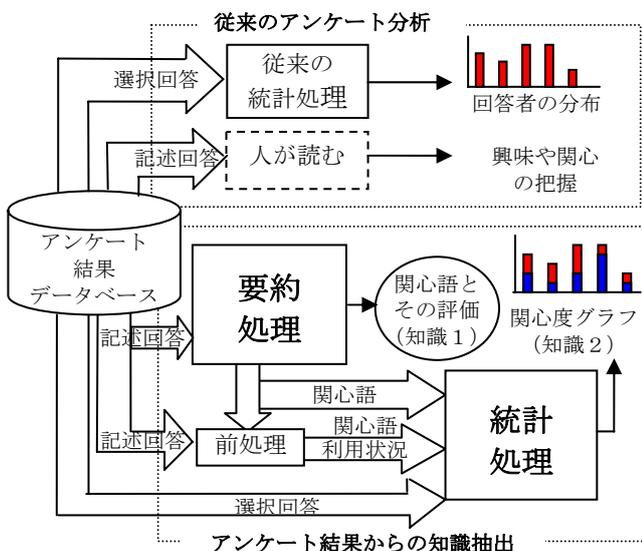


図1 アンケート分析の方法

3. アンケート結果からの知識抽出

3.1 概要

本研究では、記述回答に対して要約処理を行い、アンケート回答者の関心語とその評価を抽出し、また、選択回答などから特定の関心語に対する「関心度グラフ」を自動抽出する。

今回、本研究の検証データとして「大阪工業大学情報科学部設立10周年記念シンポジウム」のアンケート結果を用いた。以下では、要約処理と統計処理の詳細について述べる。

3.2 要約処理

要約処理は、図2のように6つの処理で構成される。すなわち、1)形態素解析により記述回答の文章を分解し、2)名詞の重要度の算出結果から、3)関心語を抽出する。また、4)記述回答の類似度の算出をするとともに、5)関心語と形態素解析から得られた動詞、形容詞により手がかり語を抽出し、6)記述回答の類似度から手がかり語の重要度を算出する。以下にその詳細を記す。

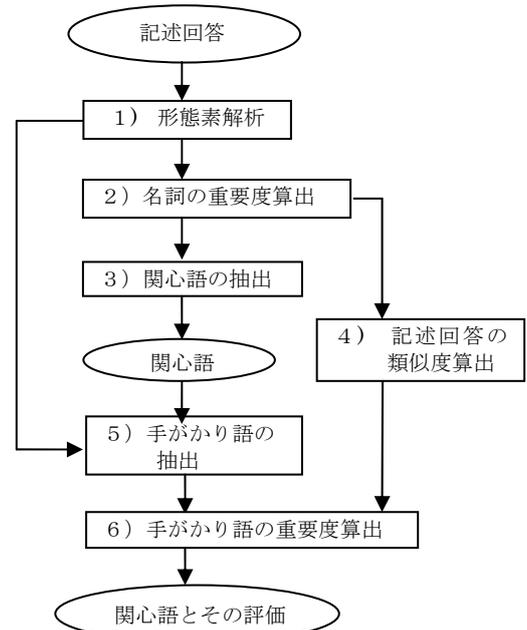


図2 要約処理の流れ

1) 形態素解析

形態素解析システム「MeCab」を用いて、記述回答の文章に含まれる用語を切り出し、各用語の品詞(名詞、動詞、形容詞など)を得る[2]。

2) 名詞の重要度算出

名詞の重要度は、その名詞の利用率と重みの積として与える。ある名詞の利用率は、全記述回答数に対するその名詞を1個以上含む記述回答数の割合である。一方、ある名

詞の重みは、一つの記述回答における全名詞数に対するある名詞の出現頻度の割合を求め、これの全記述回答における平均値とする。

3) 関心語の抽出

全記述回答における名詞の中で、最も重要度が高かったもの上位 5 つをアンケート回答者全体の関心語とする。以下の表 1 に関心語の例を示す。

表 1 関心語の例

| 順位 | 関心語 |
|-----|-----------|
| 1 位 | “ユビキタス” |
| 2 位 | “大阪工業大学” |
| 3 位 | “シンポジウム” |
| 4 位 | “インターネット” |
| 5 位 | “ネットワーク” |

4) 記述回答の類似度算出

記述回答の類似度は、二つの記述回答で一致するそれぞれの名詞の重要度を掛け合わせた合計を正規化し算出する [4][5]。

5) 手がかり語の抽出

形態素解析結果から、品詞が動詞と形容詞の用語を抽出し、関心語に対する評価を表す「手がかり語」とする。ただし、“する”、“ある”などの動詞が抽出された場合はその動詞の前につく名詞と合成し、手がかり語とする。

6) 手がかり語の重要度算出

特定の関心語を含み、類似度が一定値以上ある記述回答群を対象とし、それらの記述回答群において、関心語が存在する文から手がかり語の重要度を算出する。即ち、特定の関心語を含む文における、手がかり語の出現頻度を求め、全手がかり語数に対する割合を手がかり語の重要度とする。

特定の関心語に対する手がかり語の中で、最も重要度が高かったもの上位 5 つをアンケート回答者全体の「関心語とその評価」という知識とする。以下の表 2 に関心語とその評価の例を示す。

表 2 関心語とその評価の例

| 順位 | “ユビキタス”に対する評価 |
|-----|---------------|
| 1 位 | “おもしろい” |
| 2 位 | “楽しい” |
| 3 位 | “難しい” |
| 4 位 | “必要である” |
| 5 位 | “進歩する” |

3. 3 前処理

統計処理の前処理として、記述回答と関心語を入力し、各回答ごとに関心語を利用しているかを表す関心語利用状況を抽出する。この中で、関心語を利用している回答をその関心語の重要度の高いものと低いものに分ける。以下の表 3 に関心語利用状況の例を示す。

表 3 関心語利用状況の例

| | “ユビキタス” | “大阪工業大学” | “シンポジウム” |
|------|---------|----------|----------|
| 回答 1 | 0 | 0 | 1 |
| 回答 2 | 2 | 1 | 1 |
| 回答 3 | 1 | 0 | 2 |

2=重要度が高い, 1=重要度が低い, 0=利用なし

3. 4 統計処理

従来の統計処理により得られた選択回答者数に対し、ある特定の関心語に関心を持った（関心語の重要度が低い）回答者数と、興味を持った（関心語の重要度が高い）回答者数を示す関心度グラフを知識として出力する。以下の図 3 に関心語“ユビキタス”に関する関心度グラフを示す。

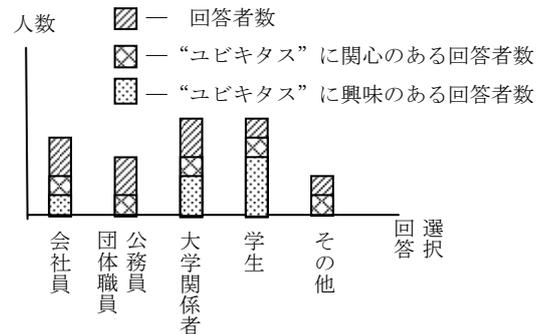


図 3 関心度グラフの例

4. まとめ

本論では、記述回答と選択回答のアンケート結果に対し、回答者の関心とその評価、および関心度グラフを生成する手法について述べた。今後、実際のアンケート結果から自動的に上記の知識抽出を行うシステムを試作し、本手法の評価を行う予定である。

謝辞

形態素解析システム「MeCab」をご提供下さった京都大学情報学研究科—日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクト、およびアンケートデータを提供していただいた大阪工業大学情報科学部に感謝いたします。

参考文献

- [1] <http://www.type-b.co.jp>
- [2] 形態素解析システム「MeCab」Version0.95
- [3] 北 研二、津田 和彦、獅々堀 正幹：“情報検索アルゴリズム”，共立出版，2002-01
- [4] 松本 裕二、馬野 元秀、瀬田 和久：“類似性に基づいたウェブページ選択インタフェース”，電子情報通信学会
- [5] 奥村 学、難波 英嗣、上田 貞子：“自動要約”，共立出版，2003-06