3ZE-06 キーワードによる知識階層を利用したクラスタリング

山本恵理子* 塩谷勇* 宮内ミナミ*
産能大学経営情報学部

1 はじめに

アンケート調査から消費者や利用者の動向を分析し，ニーズに合った商品開発や企業の意思決定への利用が一般に行なわれている．このような調査では多くの場合，アンケート作成段階で調査目的を明確にし，仮説を立ててその仮説を検証するための項目を質問中に埋め込む手法が取られる．このため，回答の分析法と予想される結論はアンケート作成段階で判明していると言っても良い．

一般には一つのアンケートに多くの質問項目が設定され，仮説検証のための誘導質問的な項目は避けられない．加えて新聞や雑誌の記事等の外的要因も考えられる．特に，複数の決められた選択肢の中から適当な項目を選ぶ質問では顕著である．そのため，アンケート作成段階で想定した分析結果以外は得にくい事が多く，逆に結果の想定が困難な場合は従来の分析方式 [4] の適用が難しい．例えば自由形式のアンケートは回答の形式や内容が多岐に渡り結果の想定が難しく，これまでの分析手法 [4] の適用ができない．従って，アンケートの結果を見ながら分析段階で仮説を立て，立証するというプロセスが得にくい．

本研究では自由記述形式のアンケートのように，事前に結果の想定が困難なテーマに対して，人工知能的な手法によって分析を行なう [5, 6]．また，通常行なわれている仮説を立てて検証するためのアンケート作成の手掛かりを見いだす事にもなる．人工知能の言葉で言うならば，知識の獲得の手掛かりを得ることができる [3]．また，従来行なわれているアンケート内容の見直しのフィードバックを行なうことができる．

本来，アンケートは回答者の意見を十分に反映できる自由記述が望ましいが，回答の自然言語文の構造と意味を取り出して分析することは回答者の背景知識まで知る必要があり，回答者数が多い場合は実質的に困難である．そこで本研究ではアンケートの設問に対して，回答者は任意個のキーワードを自由に記述できる方式とし，個々の回答者の背景知識を全体で一つの設問の関する領域の階層知識を用いて分析する．知識階層によってアンケート間の距離を定義することができ，距離に基づいた最近傍法によるクラタ分類を行なう．クラス分類の結果から階層知識の見直し (フィードバック) を行なうことができる．また，本方式の有効性を確かめるための実験を行なう．

2 アンケートの回答と概念階層

アンケートの設問は一つで，回答者が任意個複数のキーワードを列挙できると仮定する．アンケートの各回答 (この論文ではオブジェクトとも呼ぶ) は以下の形式とする．

$$e_i : w_{i,1}, .., w_{i,n_i} \quad n_i \geq 1$$

*Keyword Clustering which Knowledge Hierarchies was used for, E.Yamamoto, I.Shioya, M.Miyauchi, Sanno College

e_i はオブジェクト id で，$w_{i,j}$ はキーワードである．ここで，$e_i \neq e_j$ $(i \neq j)$, $w_{i,j} \neq w_{i,k}$ $(j \neq k)$ とする．

キーワード集合 $K = \{w_{i,j}\}$ に基づいた概念階層 C_G は根頂点が一つの非巡回有向グラフである．

$$C_G = (K, C_N, C_E, KC)$$

ここで，C_N は有限集合，$C_E \subseteq C_N \times C_N$. $c \in C_G$ を概念と言い，$w_{i,j} \in c$ $(c \in C_N)$, KC はキーワードと概念の対応を表し，$KC \subseteq K \times C_N$. ここでは，概念階層の根頂点を一つと仮定しているが，アンケートの設問が複数，アンケートの目的が複数あるとき，非巡回有向グラフとなる．$(c_x, c_y) \in C_E$ $(c_x \neq c_y)$ ならば，c_y は c_x の上位概念，c_x は c_y の下位概念と呼び，$c_x \prec c_y$ で表す．$c_x \preceq c_y$ $(c_x = c_y$ または $c_x \prec c_y)$ に関して順序関係が成立すると仮定する．

各キーワード $w \in K$ $((w, c) \in KC)$ に対応する最小の概念 c は一意に定まると仮定する．すなわち，任意の $w \in K$ に対して，必ずある $c \in C_N$ が存在して，$w \in c'$ $((w, c') \in KC)$ ならば $c \preceq c'$ である．c を w の最小概念と呼ぶ．

例 2.1 以下はアンケートの回答例である．

id	キーワード
1	産能 伊勢原 大山 石倉橋 石倉 経営情報学部 情報学科 経営学科
2	大学 伊勢原 産業能率 小田急
3	自然 田舎 図書 コンピュータ 就職

例 2.2 概念階層の例を以下に示す．

上位概念	下位概念
産能大学	キャンパス
キャンパス	伊勢原
キャンパス	自由が丘
伊勢原	伊勢原ライフ
伊勢原	大学
組織	大学

各オブジェクト $e_i : w_{i,1}, .., w_{i,n_i}$ (単に e_i とも呼ぶ) の概念は各キーワード $w_{i,j}$ を $(w, c) \in KC)$ となる最小概念 $c_{i,j}$ に置換えたものである．

$$e_i : c_{i,1}, .., c_{i,n_i}$$

ここで，$c_{i,j} = c_{i,k}$ $(j < k)$ の場合は $c_{i,k}$ が削除されている仮定する．

例 2.3 例 2.2 の概念階層を用いて，例 2.1 のアンケート回答を最小概念を用いて記述すると以下になる．

id	キーワード
1	産能大学 伊勢原 大山 石倉橋 経営情報学部 情報学科 経営学科
2	大学 伊勢原 産能大学 電車
3	自然 図書館 計算機 就職

アンケート間の類似性を距離で表す．アンケートは複数の最小の概念で特徴付けされるから，最初に概念間の距離を定義する．概念 c と c' の概念階層上の (方向を考え

ない) 最短経路パスの長さ d とするとき，関連度 $r(c,c')$ (距離) は $\frac{1}{1+d}$ とする．関連度が 1 に近いほど関連性が高く，0 に近いほど関連が薄いと考える．

各アンケートは複数の概念で特徴付けされるために，アンケート間の距離を定義する必要がある．アンケート id_1 から見たアンケート id_2 の距離を以下のように定義する．id_1 と id_2 に対応する概念を以下とする．

id	概念
id_1	$c_1,..,c_m$
id_2	$c'_1,..,c'_n$

c_i と c'_j 間の関連度が最も高い j を見つけ，その関連度を c_i から見た id_2 への距離 $d(c_i,id_2)$ とする．アンケート id_1 と id_2 の間の距離 $m(id_1,id_2)$ は $(\sum_i d(c_i,id_2)\times p(c_i)+\sum_j d(c'_j,id_1)\times p(c'_j))/2$ で定義する．$p(c)$ は概念 c の発生確率．距離は直観的には他の概念との近さの期待値である。

アンケート id_1 の重要度は他の概念との距離から $\sum_{id_2(id_1\neq id_2)} m(id_1,id_2)$ で定義する重要度の数値が大きければ大きいほど重要である．一方，小さければ，重要度が低い．

アンケートの分析は重要度の高いものを基軸にして，最近隣法で分類を行なった．

3 実験

本方式の有効性を確かめるために，産能大学経営情報学部の 1999 年度前期配当講義科目 'データベースの設計と利用 1' を履修している学生からのアンケート結果に基づいて，産能大学の学生像に対する意識調査を試みる．回答で得られるキーワードから，設問領域に関する階層知識を構築し，階層知識に基づいた最近傍法に基づいたクラタ分類を行なう．

学生に '産能大学を特徴付けるキーワード' を任意複数個列挙の上回収して得た結果に基づいて提案手法で分析を行なった．回答の概要を図 1 に示す．

件数	148
延べキーワード数	767
キーワード数	205
平均キーワード数	$\frac{767}{148}=5.18243$
第 1 キーワード数	52(148)
第 2 キーワード数	50(146)
第 3 キーワード数	71(138)
第 4 キーワード数	68(117)
第 5 キーワード数	58(88)
第 6 キーワード数	42(56)
第 7 キーワード数	23(35)
第 8 キーワード数	17(17)
第 9 キーワード数	8(9)
第 10 キーワード数	6(6)
第 11 キーワード数	2(2)
第 12 キーワード数	2(2)
第 13 キーワード数	2(2)
第 14 キーワード数	1(1)

括弧内はキーワードの延べ数.

図 1: アンケート回答の概要

各キーワードに対して，つぎの概念を割り当てた (一部)。右がキーワードであり、左が対応する概念である。

規模 2学科
大学 4年制
大学 4年制大学
伊勢原ライフ のんびり
伊勢原ライフ アメリカハナミズキ
計算機 インターネット
就職 インターンシップ
大学 カレッジ
キャンパス キャンパス
計算機 コンピュータ
計算機 コンピュータネットワーク
伊勢原ライフ サンム

回答者が列挙したキーワード中に否定的な語は無いため，概念階層作成で概念間の肯定的な上位概念，下位概念を構築すれば良い．以下のような概念階層を作成した (一部)．左がキーワード，右が割り当てが概念である．

歴史 設立
大学 経営情報学部
大学 学長
経営情報学部 経営学科
経営情報学部 情報学科
キャンパス 自由ヶ丘
キャンパス 伊勢原
伊勢原 規模
伊勢原 伊勢原ライフ
伊勢原 電話
伊勢原 自然
通学 バス
通学 電車

重要度の最もたかいものを見つけ，この近傍の存在する回答を最大 10 個抽出して一つのグループとし，この内の最大 3 個を抽出した．残りのアンケートに対して，同様の操作を繰り返した．主なものを以下である．左から，グループ番号，回答番号，残りが回答のキーワードである．

1 1 産能 伊勢原 大山 石倉橋 石倉 経営情報学部 情報学科 経営学科
1 23 経営情報学部 経営学 情報学 伊勢原市 私立大学
1 45 経営情報学部 情報学科 経営学科 伊勢原 産業能率 瑞木 瑞木祭 大山
2 123 産能 大学 伊勢原 上粕屋 経営情報
2 2 大学 伊勢原 産業能率 小田急
2 84 大学 経営情報 4年制 産能 神奈川 単科
3 141 自然 コンピュータ 図書
3 3 自然 田舎 図書 コンピュータ 就職
3 58 産業を育て能力をのばす 田舎 落ち着いている

解析の結果，類似のキーワードを列挙している回答同士がグループ化され，大学に対してどのようなイメージを抱いているかの特徴を得ることができた．

4 おわりに

キーワード形式のアンケートの解析手法を提案し、実験によってその有効性を確認した。本手法は，回答が自由形式のキーワードで記述するため，事前に結果の想定が難しい調査や事前の調査などに有効と考えられ，適用する範囲は広く，本手法は有効であると考える．また，クラス分類の結果から階層知識の見直し (フィードバック) を行なうことができる．

参考文献

[1] Y. Cai, N. Cercone and J. Han, Attribute-Oriented Induction in Relational Databases, Knowledge Discovery in Databases, MIT, 1991.

[2] 田口研治, 選択的属性指向帰納法 -情報理論に基づく新しい基準の提案-, 人工知能学会研究会資料, SIG-FAI-9403-9 (3/3), 65-71, 1995.

[3] 大須賀, 佐伯 (編), 知識の獲得と学習, オーム社, 1987.

[4] 田中豊, 脇本和昌, 多変量統計解析法, 現代数学社, 1987.

[5] Knowledge Discovery, Comm. of ACM, , Vol. 42, No. 11, 1999.

[6] Takao MIURA and Isamu SHIOYA : Mining Type Schemes in Databases, DEXA'96, 1996.