

概念グラフを用いたアンケート中の意見抽出および 可視化手法の提案

岡田, 真
大阪府立大学

廣川, 佐千男
九州大学情報基盤研究開発センター

橋本, 喜代太
大阪府立大学

<https://hdl.handle.net/2324/1498228>

出版情報 : パーソナルコンピュータ利用技術学会「情報と社会」研究会 : 第3回研究会. 1 (3), 2013
バージョン :
権利関係 :



概念グラフを用いたアンケート中の意見抽出および 可視化手法の提案

A Suggestion of a Method to Extract and Visualize Opinions in Questionnaire Surveys using a Concept Graph

岡田 真¹・廣川 佐千男²・橋本 喜代太¹

Makoto OKADA, Sachio HIROKAWA, Kiyota HASHIMOTO

1 大阪府立大学 (〒599-8531 大阪府堺市中区学園町 1-1)

2 九州大学 (〒812-8581 福岡県福岡市東区箱崎 6-10-1)

本研究では、アンケート回答者の興味や意見を可視化し取得するために、アンケート中の年齢などのカテゴリカルな情報と自由記述文中の単語から概念グラフを生成し、回答者の意見抽出および取得の方法として利用する手法を提案する。提案手法の有効性を確認するために、新規就農希望者を対象としたアンケートの調査結果を用いて概念グラフを生成し、その有効性について調査を行った。

Key words: アンケート, 意見抽出, 意見の可視化, 概念グラフ, 新規就農

1. はじめに

企業や政府や行政機関、また教育機関などにおいて、さまざまな場面や目的でアンケートを行っている。いうまでもなくアンケートにはさまざまな有益な情報が含まれており、それらに含まれる情報すべてを最大限有効に活用することはアンケート実施側には常に切望されている。

アンケート中の質問は大きく2種類に分けられる。まず“年齢”，“職業”，“満足度”など項目と何段階かに分けてその度合を選択するカテゴリカルな質問、次に設問に対して回答者が自由に記述する自由記述文中心の質問である。アンケートはこれら二種の質問を織り交ぜて構成されることが多い。

カテゴリカルな質問については、集計やそれらをもとにした数値的な分析が行いやすいため、アンケート後の報告などで情報を集約し、グラフや統計的な情報として提示されることが多い。しかし、後者の自由記述文中心の質問に関してはそれらの中に含まれる意見などを抽出し、集約されずに終わる場合がままある。その理由としては、記述文中に含まれる情報を抽出する際には記述文の内容を理解が必要とされるが、それには設問やその背景知識に対する深い知識を持つ熟練者によるチェックが必要とされ、人的な負担が大きいこと、さらにそれを自動的に十分な精度で行うのはまだ難しいことがあげられる。

だが、自由記述文中には回答者の心情や意見が多く含まれている。それらは定式化されたカテゴリカルな質問では集めることが難しく、回答者の生の意見を参照できるということからその重要性も高い。それら意見の自動抽出もしくは評価者のチェックの補助手法が存在すれば、アンケート実施者には大きな利益となると予想される。

また、意見抽出の分野ではウェブ上のレビュー記事などを対象とした意見の抽出をはじめとしたさまざまな研究がなされているが^{1),2)}、それらの多くは自由記述文中の意見抽出のみを中心としており、アンケートによくみられるようなカテゴリカルな情報と自由記述文中の情報を複合的に参照するような手法はあまり見られないのが現状である。

意見抽出関連の研究としては大規模データからの意見抽出を対象としたものが主流となるが、これらの手法は小規模のアンケート結果には用いることができない場合も多い。アンケートには数百から千回答程度の小規模のデータしか集まらないようなものも数多く、そのようなデータにはそれらを対象とした手法が求められる。

そこで、本研究では、新規就農希望者に対して行ったアンケートを用いて、概念グラフを利用したアンケート回答者の意見抽出手法についての提案を行う。以下2章で就農希望者向けのアンケートについて、3章で概念グラフの概要と就農希望者のアンケートを用いて作成した概念グラフの例について述べ、4章で今後の課題について述べる。

2. 新規就農希望者向けのアンケートの調査

新規就農希望者に対するアンケートについて説明する。

本研究で用いたアンケートは、全国農業会議所から提供された2011年と2012年に実施されたものである。全国農業会議所は農業・農業者の公的な代表機関である農業委員会系統組織の全国組織であり、農業に関する様々な取り組みを行っている。その中で新規就農者を対象としたものは「新・農業人フェア」と呼ばれる新規就農者向けのイベント、各道府県が行う新規就農に関する相談会、農業インターンシッ

ブ、就農情報ウェブサイト「あぐなび」などがある。今回、それらのイベント参加者やサイト登録者に対して全国農業会議所の協力でアンケートを行った。アンケートはそれぞれ前年の参加者などを対象に行われた。表1にその内訳を示す。

2011年と2012年に行われたアンケートは質問項目が違うが、カテゴリカルな質問項目と自由記述中心の質問項目が含まれている。本稿では主に2012年のアンケートを中心に説明する。

表1 新規就農希望者向けアンケート結果

	2011年度 (人)	2010年度 (人)
総数	1117	325
農業法人への就職就農、独立就農	114	17
就農への研修中	143	51
農業体験	151	54
情報収集	381	181
その他	325	22

2012年のアンケートではまず“回答日時”、“相談の経緯”、“年齢”、“性別”、“現在住む都道府県”、“現在の就農状況”の6つのカテゴリカルな質問を行い、アンケート実施時の就農状況の結果に合わせて5つのグループに分けて、それぞれに合わせてカテゴリカルな質問と自由記述で構成されているアンケートを行い、集計を行った。

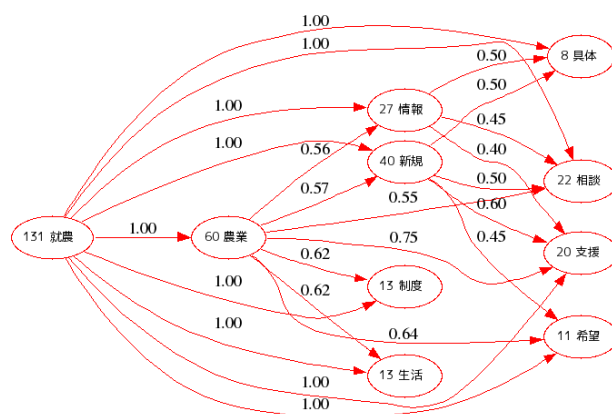
3. 概念グラフ

分類は対象を分析する最も基本的な手法である。どのような属性に着目して分類を行なうかにより得られる結果は異なる。属性集合と対象集合が互いに特徴付けとなるような組を概念とよび、対象全体の分類階層を束として表現したものが概念束である^{3),4)}。本稿では、文集合を対象とし、それらの文に現れる単語を属性とする概念束を考える。概念束では、一つの概念（ノード）は文集合と単語集合の組として捉えられ、互いに特徴づけあっている。また、隣接する概念により概念の分類を行っている。対象文書群に現れる全ての単語の個数を N とすると、可能な単語集合は N の冪乗となり膨大な数になり、全体を可視化することは現実的にはできない。本稿では、引用文献^{5)~7)}で提案されている概念グラフを用いる。

概念グラフは以下のように定式化できる。文書集合 D 、単語集合 W 、および D, W の関係を表す $M \subseteq D \times W$ をまとめて (D, W, M) と書き文脈と呼ぶ。集合 D の要素 d 、集合 W の要素 w について文書 d に単語 w を持っているとき $(d, w) \in M$ と表記する。 $X \subseteq D, Y \subseteq W$ のとき (X, Y) が $\text{doc}(Y)=X$ かつ $\text{word}(X)=Y$ となるならば、 (X, Y) を概念と呼ぶ。このとき (D, W, M) の概念全体を $\text{CL}(D, W)$ と書く。ただし、 $\text{doc}(Y)=d \in D, \forall w \in Y d$ は w を含む、 $\text{word}(X)=w \in W, \forall d \in X d$ は w を含む。 $\text{CL}(D, W)$ の要素 $(X, Y), (X', Y')$ に対し、 $Y > Y' \Leftrightarrow \text{doc}(Y) \supseteq \text{doc}(Y')$

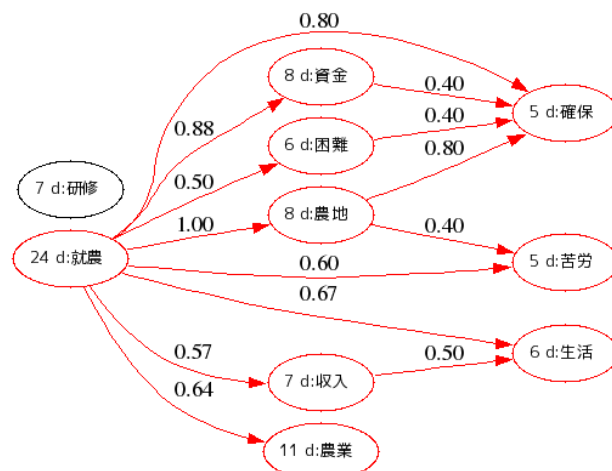
と定義すると、順序集合 $(\text{CL}(D, W), >)$ は束となり、これを概念束という。二つの単語 $u, v \in W$ に対し、 $u > v \Leftrightarrow \text{doc}(u) \supseteq \text{doc}(v)$ により定義される順序集合 $(W, > \text{CN})$ を (D, W) の分類ネットワークと呼び、 $\text{CN}(D, W)$ と表す。閾値 α を使って、 $\text{doc}(u) \supseteq \text{doc}(v)$ の条件を、 $|\text{doc}(u) * \text{doc}(v)| \geq \alpha * |\text{doc}(v)|$ と一般化した有向グラフを参考文献⁶⁾では概念グラフ (Concept Graph) とよんでいる。分類ネットは閾値が $\alpha=1$ の概念グラフといえる。

図1は、「就農」をキーワードとして得られる131件の検索結果に現れる頻度上位10位の特徴語に対し、閾値を0.4とした概念グラフである。情報収集しているときに相談が必要とか、新規に始めるための支援などが回答されていることが推測できる。図2は、検索語は同じだが、特徴語は問題点の部分に現れている単語に限定している。農地や資金の確保についての苦勞とか、収入や生活についての不安が現れている。特徴語として世代も加えたのが図3である。収入や生活のことを気にしているのは20代で、農地の確保を気にしているのは40代、資金を気にしているのは30代、40代であることが分る。



query="就農" 131 db=agriNV alpha=0.40 lf=1

図1 「就農」に対する概念グラフ (対象は単語, $\alpha=0.4$)



query="就農" 131 db=agriNV alpha=0.40 lf=1

図2 「就農」に対する概念グラフ (対象は問題点, $\alpha=0.4$)

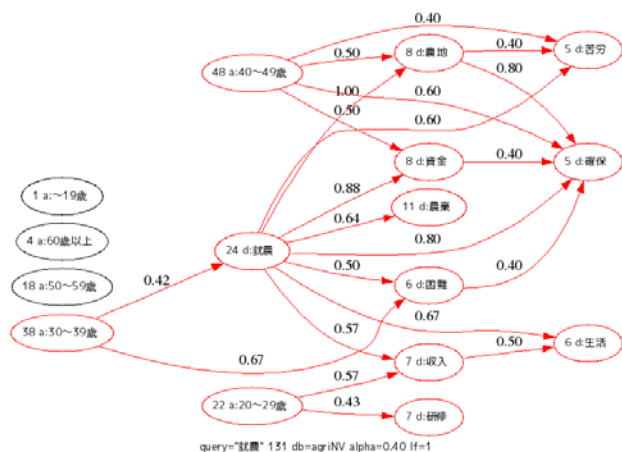


図3 「就農」に対する概念グラフ（対象は問題点と世代， $\alpha = 0.4$ ）

参考文献

- 1) 唯野良介, 嶋田和孝, 遠藤勉: 意見の重要度と客観的補足情報を考慮したレビュー要約, 言語処理学会第17回年次大会, pp.204-207, 2011.
- 2) 河合剛巨, 岡嶋 穰, 中澤 聡: Web 文書の時系列分析に基づく意見変化イベントの抽出, 言語処理学会第17回年次大会, pp.264-267, 2011.
- 3) C. Carpineto, G. Romano, Concept Data Analysis Theory and Application, John Wiley and Sons, 2004.
- 4) G. Ganter, R. Wille, C. Franzke, Formal Concept Analysis Mathematical Foundation, Springer, 1999.
- 5) Sachio Hirokawa, Takahiro Baba, Tetsuya Nakatoh, Search and Analysis of Bankruptcy Cause by Classification Network, in: L. Ballatrecche and F. Mota Pinto (Eds.): Proc. 1st International Conference on Model & Data Engineering MEDI2011, LNCS 6918, pp.152-161, 2011.
- 6) Y. Shimoji, T. Wada, S. Hirokawa, Dynamic Thesaurus Construction from English-Japanese Dictionary, Proc. The Second International Conference on Complex, Intelligent and Software Intensive Systems, pp.918-923, 2008.
- 7) T. Baba, L. Liu, S. Hirokawa, Formal Concept Analysis of Medical Incident Reports, Proc. KES2010, Vol.6278, pp.207-214, 2010.