論　文　***Article***

自由フォーマットテキスト文に対する統計的分析

原稿受付　2023 年 7 月 31 日
ものつくり大学紀要　第 13 号　(2023)　　7～11

佐久田茂

ものつくり大学　技能工芸学部　情報メカトロニクス学科

概要　　2021 年度総合機械学科インターンシップ評価票中の「研修学生の実習態度や仕事（課題）への取り組み方などお気付きの点をお聞かせください」に関して非構造テキスト文を英訳してデータを作成後、統計分析ソフト JMP のテキストエクスプローラ機能を用いて統計分析した。トピック分析、感情分析、単語選択分析等によりアンケート分析の見通しを得ることができた。

キーワード：統計，分析，テキスト文，アンケート，情報

Statistical analysis of text data in questionnaire forms

Shigeru SAKUTA

Dept. of Manufacturing Technologists, Institute of Technologists

Abstract　There are many questionnaires collected from corporations regarding internship or recruit research meeting, high school students regarding open campus, university students regarding classes and so on in the Institute of Technologists. Unfortunately, they have not been analyzed objectively and quantitatively and actually they have not been made the most use of. JMP, statistical discovery software, deals with text data by “Text Explorer” function based on a kind of statistics principal component analysis. In this paper, five kinds of visual statistical data analyses, that is, term/phrase analysis, singular value decomposition, topic analysis, sentiment analysis and term selection analysis, have been shown and discussed from viewpoint of its feasibility.

Key Words : statistical analysis, text, questionnaire, information

1. はじめに

現状，オープンキャンパス・インターンシップや企業説明会などでの各種アンケート，特に自由記述欄データが単年での主観的総括にしか活かされていない．もったいない．統計ソフトウエア JMP（JMP16，SAS Institute Japan 株式会社）では非構造化テキストデータ，すなわちアンケートの自由記述欄などの自由フォーマットテキスト文の統計的分析が可能である．今回，大学保有の自由記述アンケートの統計分析を試みた．

2. 分析方法

自由フォーマットテキスト文を JMP16 のテキストエクスプローラ機能にかけて分析した．テキストエクスプローラは自然言語処理（NLP）中の特異値分解アルゴリズムを使ってテキストを数値データに変換，そのデータを統計分析にかけている [1]．分析対象は，インターンシップ評価票での「研修

学生の実習態度や仕事（課題）への取り組み方などお気付きの点をお聞かせください」における，企業様からの回答テキスト文とした．理由は，甘口・辛口様々な忌憚のない意見が一番期待できると考えたからである．本報ではテキスト分析試行の意味合いを込めて，2021 年度総合機械学科学生向け（2 年生約 120 人，および 3 年生約 80 人）の企業回答に限定して分析を行なった．企業様からのコメントから１行に１文としたデータテーブルを作成した．JMP データテーブルにおけるテキスト列の１セルが１文書に対応している．なおテキストエクスプローラ中の一部の分析の対象言語が英語のみでの対応となっているので(2023/7 時点)，今回はアンケート文を英訳して分析にかけた．表１にデータテーブル（抜粋）を示す．なおアンケート文の英文への翻訳は全て著者が行なった．

図１は，テキストエクスプローラのプラットフォームである．「列の選択」で表１の「Comments」の列を選択し，「選択した列に役割を割り当てる」中の「テキスト列」にコピーして分析が進められる．

Table. 1　Data sample

No.	Grade	Comments
1	Sophomore	OK for manners as trainee. Positive thinking for Internship and for oneself were seen sometimes.
2	Sophomore	We can make him feel the principle of manufacturing through all processes in our plant.
3	Sophomore	We hope this experience makes him grow up considering his future dream as a employee.
4	Junior	He had been very serious and asked questions and reported to employee politely.
5	Junior	He tended to take time to start to ask questions.
6	Junior	If he accumulated questions before asking, I wanted him ask at once ,so I told him asked questions once he had them.

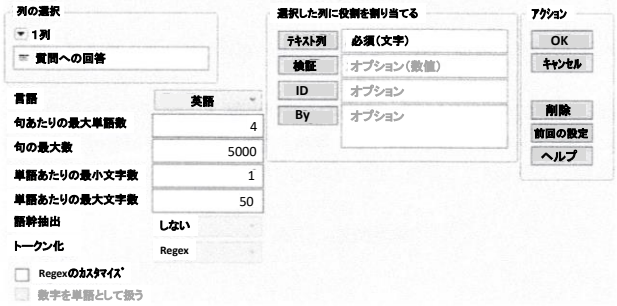

Fig.1　Platform of text explorer

3. 分析結果

3.1　出現単語・出現句の傾向分析

図２は単語と句の出現結果である．度数分布表の形でアンケート文中での出現単語・出現句の傾向を見ることができる．出現単語・出現句の傾向を見ながら，ケースに応じて文書間の違いを読み取るのに役立たない単語を分析から除いていく．具体的には単語を選択して右クリックして，“ストップワード”に追加する．

例えば，旅客機の墜落原因アンケートでは，「pilot」を含む単語は他の単語より頻繁に出現してくるが，文書間の違いを読み取るためには役立たないため，ストップワードに指定して分析から除く [2]．本報ではストップワード無しでの分析で不具合が見られなかったので，ストップワードを特に指定せずに分析を進めた．

単語と句のリスト

単語	度数
work·	211
good	121
think·	88
ask·	68
internship·	66
so	64
best	53
posit·	51
question·	47
very	47
thing·	45
greet·	44
use·	43
communic·	42
employe·	42
engin·	42
seem·	42

句	度数
good engineer	24
think he should	21
asked questions	18
very much	18
so i think	10
didn t	9
good engineer if	8
little by little	8
engineer if	8
given problems	8
many things	8
quick learner	7
communicate with others	6
couldn t	6
too much	6
unclear things	6
seemed to be quiet	5

Fig. 2　Term/phrase analysis

3.2　特異値分解

特異値分解 [3]によってテキストデータを数値データに変換する．その変換された数値データに基づいて統計分析が実施される．特異ベクトルの分析でその次元（特異値分解プロット上の X 軸・Y 軸）が何を表しているかがわかる．特異値分解プロットの各点は，文書または単語と紐付けされていて，プロット上の点を選択（クリック）することで該当する文書や単語を表示することができる．

図 3 は表 1 のデータの特異値分解である．2 年生と 3 年生には大きな差異はないことがわかる．また分布に大きな偏りがないことから，3.3 および 3.4 にて後述のトピック分析・感情分析に移行する．偏りがある場合は，縦軸・横軸がどういう特性をもった軸なのかを個別の文書内容を参照して推測すると理解が進むことがある．なお特異値分解は，トピック分析を行う前のステップとしても位置付けられる．

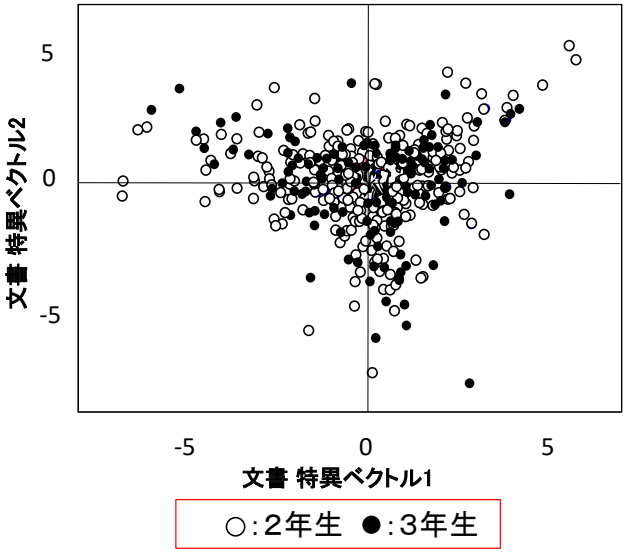

Fig 3　Singular value decomposition

3.3　トピック分析

テキストエクスプローラでアンケート文中での出現単語・出現句の傾向を把握し，特異値分解後にトピック分析 [4]に移ることができる．テキストの内容を解釈しやすくするため，文書単語行列の特異値分解した結果を Varimax 回転し.単語を「トピック」と呼ばれるグループにまとめる．アンケートの代表的トピック上位（例えば上位 5 トピックや上位 10 トピックなど:トピックの数は変更可能）を示唆できる．図 4 はトピック分析結果である.最上位のトピック（図 4 中のトピック 1）は，“ask”，“question”，“understand”，“positively”などの負荷量の大きな単語より「不明な点を積極的に質問して欲しい／していた」，

２番目のトピックは「将来このインターンシップが役立つことを希望している」，３番目は「今の学修を継続すれば将来は良いエンジニアになれると思う」などであったことが推測できる．トピック 1 の“ask”，“question”の負荷量がそれぞれ約 0.69，0.68 と他の負荷量（0.38～）に比べて顕著に大きな値となっていることから，全アンケート中で「質問する（ask a question）」に関してのトピックが突出していることが分かる．

トピック 別上位負荷量

トピック１		トピック２		トピック３	
単語	負荷量	単語	負荷量	単語	負荷量
ask·	0.6902	futur·	0.4801	if·	0.5437
question·	0.6770	experi·	0.4781	engin·	0.4695
unclear·	0.3831	use·	0.4774	think·	0.4340
thing·	0.3761	hope·	0.4097	good	0.3763
posit·	0.3062	make·	0.3666	continu·	0.3478
understand·	0.2977	internship·	0.3409	much	-0.3263
greet·	-0.2814	join·	0.3294	studi·	0.3117
many	0.2761	univers·	0.2463	measur·	-0.2855
t·	0.2495	knowledg·	0.2459	more	0.2733
manner·	-0.2277	without	-0.2112	problems	-0.2626
didn't	0.2245	report·	0.2077	result·	-0.2594
		greet·	-0.2077	very	-0.2443

Fig 4　Topic analysis

3.4　感情分析

テキストエクスプローラでアンケート文中での出現単語・出現句の傾向を把握した後，感情分析 [5]に移ることができる．感情分析は，辞書に基づいて文書内の感情語を特定し，それらの語に対して，肯定的・否定的・全体的なスコアを付けることができる．簡単に言うと感情分析で総体的感情の度数分布を見ることができる．なお現状，JMP の感情分析は，英語のみしか扱えない [6]．

図 5(a)(b)は感情分析の結果である．図 5(a)の「文書のスコアと総体的感情」を見ると，アンケート記述が文書スコアでプラスとなっていて，総じて肯定的に記されていることが定量的・客観的にわかる．

スコア 87.5，62.5，37.5，－62.5 の度数をそれぞれ 75，150，25，25 として，プラス感情総計とマイナス感情総計を簡易的に[スコア]×[度数]より計算すると，

プラス感情総計：

87.5×75＋62.5×150＋37.5×25=16875

マイナス感情総計：

62.5×25=1562.5

すなわち,

プラス感情総計／マイナス感情総
=16875/1562.5≒11

となり,プラス感情がマイナス感情の約 11 倍を占めることが分かる.

各文書と全体的スコアは図上で紐づいていて,例えば図 5(b)では文書 624 の感情スコアが“96 点”と高く,その一方文書 617 のスコアが“−60 点”と低いことが分かる.図 5(b)中の“合計”は文書中での肯定・否定的のスコアの総計,“平均”は合計値を文書中の肯定・否定的語句の出現数で割った数である.ちなみに文書 624 は,

“I think he was very good at understanding our instructions and advice.”

(われわれの指示・アドバイスを良く聞いてくれた)

文書 617 は,

“He sometimes took a nap, which shows problems of our way asking works.”

(時々居眠りしていて,われわれの教え方が悪いのかなとも思った)

であった.

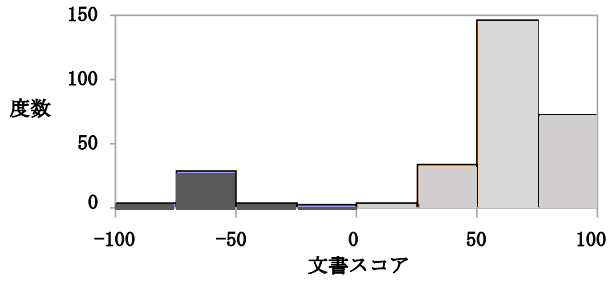

Fig 5 (a) Sentiment analysis :Chart

文書	肯定的 合計	肯定的 平均	否定的 合計	否定的 平均	全体スコア
616	60	60	0	0	60
617	0	0	-60	-60	-60
618	0	0	0	0	0
619	0	0	0	0	0
620	60	60	0	0	60
621	0	0	0	0	0
622	60	60	0	0	60
623	90	90	0	0	90
624	96	96	0	0	96
625	150	75	0	0	75

Fig 5 (b) Sentiment analysis :Scores

3.5 単語選択分析

テキストエクスプローラでアンケート文中での出現単語・出現句の傾向を把握した後,単語選択分析[7]に移ることができる.

単語選択分析を用いると特定の応答変数を最もよく説明する単語を特定できる.例えば仮にオープンキャンパス来場者とアンケート結果を紐付けできるとすると,アンケートで「後日ものつくり大学に入学した(Yes)か,否(No)」か」で,Yes を最もよく説明する単語(例えば,実技,装置,インターンシップ,就職率,学生プロジェクト,etc)や,No を最もよく説明する単語(例えば,授業料,通学時間,立地,生活費,博士,etc)をアンケート結果から統計的に見出すことができる.

4. まとめ

インターンシップ評価票中の研修学生について受入れ企業様からの気付きに関する自由フォーマットテキスト文に対して統計的分析を試行した.

1) トピック分析を用いてアンケートの概要の上位数パターンを客観的かつ定量的に求めることができた
2) 感情分析を用いてアンケート総体のプラス／マイナスの感情傾向を客観的かつ定量的に求めることができた

また単語選択分析を用いて様々な結果に影響を及ぼすキーワードを抽出できるため,特定の応答変数(興味ある結果)とアンケート内容が紐付けられれば,単語選択分析は施策立案に有効な分析手法になると考えられる.

今回の試行を踏まえて今後は,以下の取組みを進める予定である.

1) インターンシップ評価票の層別分析
   ・年度毎
   ・学科毎
   ・業種毎
2) オープンキャンパスや授業アンケート等の各種アンケート結果の分析,例えば,
   ・入学者／未入学者のオープンキャンパスアンケート結果の差異調査

・インターンシップの学年の違いによる企業評価差異調査
・就職実績有無による企業アンケート結果の差異調査
・就活に関する企業アンケートの変遷調査

さらにテキストエクスプローラは感情分析機能を使わなければ，日本語にも対応しているので，今後英語版と日本語版の分析比較も実施したいと考えている．

謝辞

本研究は，「2022 年度ものつくり大学教育力・研究力強化プロジェクト」の採択を受け支援を頂きました．またデータ整理には Dewahewage Dulitha Pasindu 氏の協力を仰ぎました．感謝申し上げます．

文献

1) https://www.jmp.com/support/help/ja/16.2/index.shtml#page/jmp/text-explorer.shtml, (参照 2023-03-16).
2) SAS Institute Inc., JMP16 ドキュメンテーションライブラリ,2021, p1409.
3) SAS Institute Inc., JMP16 ドキュメンテーションライブラリ,2021, p1392.
4) https://www.jmp.com/support/help/ja/16.2/index.shtml#page/jmp/topic-analysis.shtml, (参照 2023-03-16).
5) https://community.jmp.com/t5/JMP-Blog/Sentiment-Analysis-comes-to-JMP-Pro-16/ba-p/315123, (参照 2023-03-16).
6) https://www.jmp.com/support/help/ja/16.2/index.shtml#page/jmp/sentiment-analysis.shtml, (参照 2023-03-16).
7) SAS Institute Inc., JMP16 ドキュメンテーションライブ