データ分析における目的設定の有無による分析過程の違い
Differences in analytical process caused by purpose setting in data analysis

辻　泰輝[†]，山崎 治[†]
Taiki Tsuji, Osamu Yamazaki

[†] 千葉工業大学情報科学部
Faculty of Information and Computer Science, Chiba Institute of Technology
s1532102hu@s.chibakoudai.jp

概要

The purpose of this study is to investigate the effect of prior purpose setting on the analytical process. In the experiment, participants were divided into two groups, "group with purpose / group without purpose", and asked to analyze data of sales of virtual stores using multiple graphs. As a result of protocol analysis, the difference appeared in the way of viewing and using the graphs. Participants in groups with purpose viewed more composite graphs and focused on important elements related to the purpose.

Keywords : data analysis, analytical process, problem definition

1. はじめに

ビッグデータを企業の意思決定の材料として活用するため，ビジネス現場でのデータ分析や，データサイエンティスト育成への期待が高まっている．データ分析とは抱えている問題を解決するための手段の一つであり，分析の目的を明確にしなければ大量のデータから本当に必要な知見を得ることはできない．

データ分析のプロセスとして藤本・青山[1]はデータ駆動要求工学 D2RE（Data-Driven Requirements Engineering）の枠組みと，D2RE に基づく「A*プロセス」を提案した（図 1）．

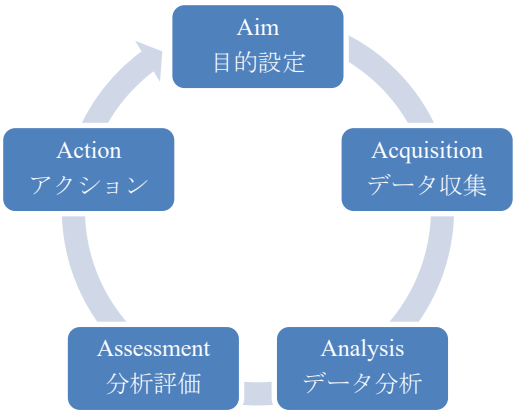

図 1　A*プロセス図（[1]を改変）

「A*」とは Aim（目的），Acquisition（データ収集），Analysis（分析），Assessment（評価），Action（アクション）の 5 つのアクティビティの頭文字を表している．

またデータ分析は目的を設定する者と実際に分析を行う者が同一であるとは限らない．企業経営者が保有しているデータを活かしたいと考え，データサイエンティストに分析を依頼するというような状況も考えられる．そのような状況下で分析を依頼する側と実際に分析を行う側で目的の乖離があった場合にも目的設定の不明確さによって必要な分析結果を導き出せない可能性がある．

本研究では A*プロセスにおける Aim（目的設定）を重視し，その目的が分析者の思考や着眼点にどのような影響を及ぼすのかを調査するとともに，目的設定の重要性を明らかにするための実験を行う．

ただし，本研究においては目的の有無が分析者の「思考や着眼点」に与える影響の調査を目的としているため，データの選択，加工，統計処理等は行わせず，あらかじめ実験者が用意したグラフを読み取る行為を「分析」とする．その中でも読み取った内容の解釈（思考）や着目するグラフ（着眼点）に違いが現れるのではないか，という仮定の元で実験を行う．

2. 目的

本研究では，複数のデータを分析することで有意義な事実を発見し，予測を立てる活動に注目し，「目的設定」を持ってデータ分析を行った場合と，そうでない場合での分析過程やその結果に差が現れるのかを調査する．この調査を通じて，データ分析活動における「目的設定」の重要性を明らかにすることを本実験の目的とする．

3. 方法

3.1 実験参加者

情報科学もしくは経営工学を専攻とする大学 4 年生 12 名が 2 人 1 組ごとに実験に参加した． 計 6 組の参加者を，目的なし群（3 組），目的あり群（3 組）に分けた．

3.2 実験計画

1 要因 2 水準参加者間計画で行う．独立変数として目的の有無を取り上げ，「目的を与える」／「目的を与えない」の 2 水準を設ける．

3.3 材料

Kaggle 社の Web ページに掲載されている「Store Item Demand Forecasting Challenge（ https://www.kaggle.com/c/demand-forecasting-kernels-only/data)」という商品売上のデータを用いた．同データを加工し，仮想的な実験用データとした．実験用データは「月」「日」「曜日」「店舗」「アイテム」「年齢」の 6 項目で構成され，それぞれに対して「売上（個数)」が示される．また資料として， 1 つの項目に対する売上を示すグラフ（単純集計）を 6 個，2 つの項目を掛け合わせたものに対する売上を示すグラフ（クロス集計）を 15 個 Excel 上で作成した．

3.4 手続き

実験は「データの観察・話し合い（25 分間)」・「記述（制限時間なし)」の二段階で構成される．目的あり群，目的なし群の共通目的として，実験用データから読み取れる「現状」と，それに対する「改善案」を提出することを求めた．この際，目的あり群のみに「既存商品・既存店舗についての弱みを知りたい」という分析の目的を伝えた．またグラフ閲覧・操作の過程を記録するために，スクリーンキャプチャソフトで録画した．さらに，参加者 2 人による協調活動の過程を記録するために，ビデオカメラ(SONY 株式会社•HDR-CX120)で参加者の様子を録画し，iPhoneXS で参加者の発話音声を録音した．

4. 結果

4.1. グラフ閲覧の様子に関する分析

4.1.1 単純集計とクロス集計の閲覧率

各群が課題中に閲覧したグラフが単純集計であったのか，それともクロス集計であったのかについて，割合では，図 2 に示す．

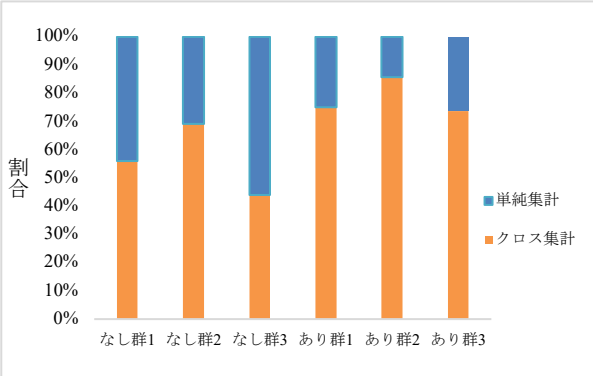

図 2 単純集計とクロス集計の閲覧率

本実験では分析資料として単純集計グラフを 6 個，クロス集計を 15 個作成した為，おおよそ単純集計 30%，クロス集計 70%の割合であった．図 2 から目的なし群は 70%以下，目的あり群は 70%以上クロス集計を閲覧した結果となった．

4.1.2 グラフごとの閲覧数（平均）

また，目的なし群，目的あり群別にグラフ 21 個ごとの閲覧数を図 3 に示す．縦軸が閲覧数の平均値，横軸がグラフの番号を示す．青色の棒グラフは目的なし群，オレンジ色は目的あり群を示し，グラフ番号の 1〜6 番が単純集計のグラフ，7〜21 番がクロス集計のグラフである．

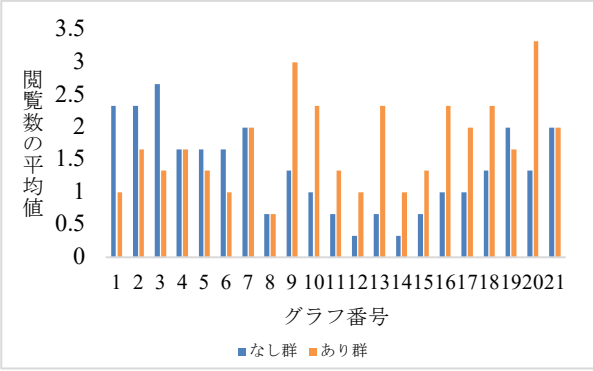

図 3 グラフごとの閲覧数（平均）

目的なし群はクロス集計よりも単純集計のグラフを，目的あり群は単純集計よりもクロス集計のグラフを閲覧していることが図 2 からも分かる．

目的なし群と目的あり群との間で，閲覧数の差が特に大きいグラフは 9（店舗×月)，10（アイテム×月)，13（店舗×日)，16（店舗×曜日)，20 番（店舗×年齢)のグラフである．全てのグラフがクロス集計であり，「アイテム」もしくは「店舗」の項目が含まれるグラフである．

4.1.3 グラフ遷移の連続箇所

各群の分析の過程で，閲覧されたグラフの順序に注目した分析を行った．3 回以上連続で同じ項目を含むグラフを遷移した場合，その箇所を遷移の連続箇所とした．この際，単純集計とクロス集計の区別はしていない．グラフ遷移の連続箇所を視覚化することで，2 人 1 組の参加者がどの項目を軸にデータとグラフを閲覧していたのかを観察出来ると考えた．各 6 組の実験のグラフ遷移の連続箇所をまとめた表とグラフが表 1 と図 4 である．

表 1 の（ ）内の数値は同じ項目が連続した回数を表す．例として，「店舗」→「店舗×アイテム」→「年齢×店舗」の順番でグラフが閲覧された場合，連続箇所として 1 カウントされる．その合計値が表 1 の一番下の連続箇所数として記されている．またこの際「店舗」という項目が 3 回連続で続いている為，「店舗（3）」と表記される

表 1 グラフ遷移の連続箇所

	なし群 1	なし群 2	なし群 3	あり群 1	あり群 2	あり群 3
	店舗（3）	店舗（5）	アイテム（3）	店舗（3）	日（4）	店舗（8）
	アイテム（4）	アイテム（3）		月（3）	店舗（4）	曜日（6）
	年齢（5）			曜日（4）	年齢（5）	年齢（3）
	曜日（4）			店舗（3）	アイテム（3）	年齢（4）
					年齢（3）	店舗（4）
					アイテム（3）	

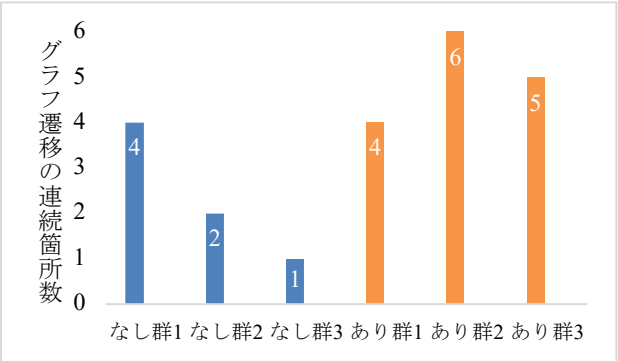

図 4 グラフ遷移の連続箇所数

目的なし群は 4 箇所以下，目的あり群は 4 箇所以上グラフ遷移の連続が見られた．各 6 組が連続して着目する項目に傾向は現れなかった．

4.2. 発話内容に関する分析

4.2.1 発話内容の分類

録音した発話を書き起こし，その発話内容を「計画」，「グラフから読み取れる内容」，「独自の解釈」の 3 つのグループに分類した．「計画」は，次に閲覧するグラフを決める様な発話，「グラフから読み取れる内容」は，グラフが示すデータに関する発話で，人による解釈の違いが現れない発話，「独自の解釈」はグラフが示すデータに関する解釈や，原因推測（売上），また「月」「日」「曜日」「店舗」「アイテム」「年齢」以外のグラフに示されていないデータを用いた解釈や原因推測（売上）の発話，と定義した．またその発話例を表 2 に示す．

表 2 3 つの分類（例）

3 つの分類	例
計画	店舗のグラフを見よう
グラフから読み取れる内容	店舗 1 と 5 の売上が低い
独自の解釈	店舗 1 と 5 は立地が悪いんじゃない？

各組の発話における 3 つの分類の割合を表したグラフが図 5 である．

表 2 と同様に「計画」：オレンジ，「グラフから読み取れる内容」：緑，「独自の解釈」：青の 3 色で表されている．また A，B は各群 2 人の参加者を指す．

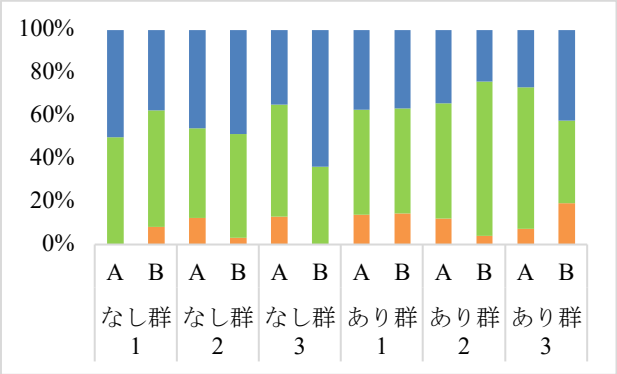

図 5　3 つの分類の割合

まず「グラフから読み取れる内容（緑）」と「独自の解釈（青）」について，目的あり群は「グラフから読み取れる内容」が，目的なし群は「独自の解釈」が割合を多く占める結果となった．

次に「計画（オレンジ）」について，目的あり群は A，B の 2 人から発せられているのに対し，目的なし群においてはどちらか一方に偏っている様子が読み取れた．

4.2.1 独自の解釈を分類

「独自の解釈」に分類された発話は参加者それぞれの考えが反映されている発話である．この発話内容をさらに細分化することで，参加者ごとの意図や分析の指針が読み取りやすくなるのではないかと考え「閲覧したグラフに対する評価」，「問題提起（売上），「原因推測（売上）」，「その他」の 4 つのグループに分類した．「閲覧したグラフに対する評価」は，あるグラフを閲覧した後に，そのグラフに対して観察する価値があるかないかの判断する様な発話，「問題提起（売上）」は，売上げの低い店舗や商品について，なぜ低いのかを検討する前に，「売上を上げるにはどうしたら良いのか？」という観点で改善策を考えている発話，「原因推測（売上）」は，売上が低いデータもしくは高いデータに対して，その原因を考察している発話，「その他」は，上記 3 のグループに当てはまらない発話，と定義した．またその発話例を表 3 に示す．

表 3　4 つの分類

4 つの分類	例
閲覧したグラフに対する評価	このグラフは重要そう （見る必要がなさそう）
問題提起（売上）	・1 月 2 月どうやって伸ばすか ・60 代を上げたいよね ・売上を上げるにはどうすればいいんだ
原因推測（売上）	・給料日だからじゃない？ （月末の売上が高いことに対して） ・1,5,16 が売れないのは 4,50 代が買わないからじゃない？
その他	・50 代がやたら伸びてるとことが気になるな ・土日に売れないアイテムとかあるの？

4 つの分類の割合を表したグラフが図 6 である．表 3 と同様に「閲覧したグラフに対する評価」：ピンク，「問題提起（売上）」：紫，「原因推測（売上）」：青，「その他」：白の 4 色で表されている．A，B は各群 2 人の参加者を指す．

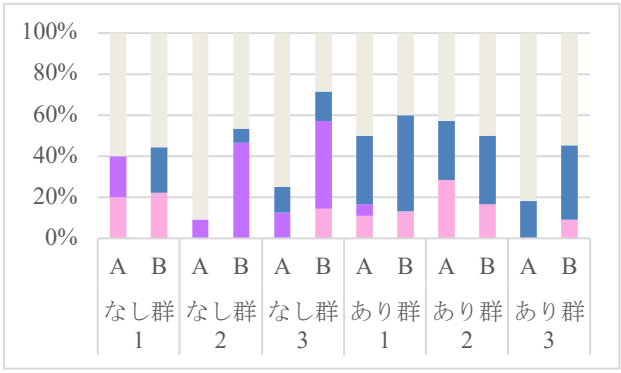

図 6　4 つの分類の割合

「その他（白）」に分類された発話を除いて目的あり群では「閲覧したグラフに対する評価（ピンク）」と「原因推測（青）」が，目的なし群では「問題提起（紫）」の発話の割合が増えるという結果となった．

5.　考察

まず，単純集計とクロス集計の閲覧率について，目的あり群の方が目的なし群よりクロス集計のグラフを

多く閲覧していることがわかる．目的あり群は「既存商品・既存店舗についての弱みを知りたい」という分析の目的を与えられていた為に「アイテム」と「店舗」の項目に着目しやすい．その結果，「アイテム」と「店舗」を他の項目と照らし合わせる為に，クロス集計のグラフをより多く閲覧したのではないかと考えられる．

次に，グラフ遷移の連続箇所について，単純集計とクロス集計の閲覧率と同様に目的あり群は，「アイテム」と「店舗」を軸に他の項目と比較しながらグラフを閲覧すると考えた．その結果，「アイテム」「店舗」の 2 項目を主としたグラフ遷移の連続箇所が多くなるのではないかと考えた．連続箇所総数は目的あり群の方が多いが，特定の項目を軸にグラフを閲覧している様子はなかった．原因として，使用した実験用データが簡略であったと考えられる．分析資料として作成したグラフは，どれも右肩上がりのグラフや，関係性は変わらず総量が増減しただけのような一定の傾向しか持たないグラフであった「既存商品・既存店舗についての弱みを知りたい」という分析目的の下，目的あり群がアイテムや店舗のデータを閲覧しても，データの簡略さ故に，「アイテムや店舗の弱み」と捉えられるような要因を見つけることが困難であったと考えられる．

最後に，参加者同士の発話内容について，「グラフから読み取れる内容」の発話が多かった目的あり群は，分析目的であった「商品と店舗の弱み」を把握するために，グラフに示される現状を整理しようする傾向にあったのではないかと考える．他方，「現状と改善案の提出」という指示しかされていない目的なし群は，閲覧したデータに対して，自身の経験等から売上の低い原因を予測し改善案を考えようした結果，グラフに示されているデータ以外の要素を加えた解釈や原因推測（売上）に関する発話が増えたのではないかと考える．「計画」に関する発話については，目的あり群は目的を与えられていたために，共通する指針に対する 2 人（A, B）の意見が出やすかったのではないかと考える．しかし目的なし群は A，B それぞれの指針が統一されておらず，2 人の内のどちらかの考えにもう片方の参加者が合わせたために，「計画」発話が偏ってしまったのではないかと考える．

また「独自の解釈」の分類について，両群とも「現状と改善案」の提出を共通の目的としていたが，目的あり群は「弱みを知りたい」という分析目的を与えられたことによって，現状を把握することを主目的として，データの観察や考察を行ったのではないかと考える．その結果「閲覧したグラフに対する評価」や「原因推測（売上）」に関する発話が増えたのではないかと考える．他方，目的なし群は改善案を出すことを主目的とし，売上の低い店舗や商品，年代をどう上げるのかのみを考えた．その為，現状を整理する様な観察を行うことやデータに対する評価基準が無かったのではなかと考えられる．結果，目的あり群の様な発話傾向は現れず「問題提起（売上）」に分類される様な発話が増加したのではないかと考える．

6. まとめ

本研究の目的は目的設定の有無がデータ分析の過程や結果の洞察に及ぼす影響を調査することであった．実施した実験においては目的設定に焦点を当て，「目的あり群／目的なし群」に分けた参加者にデータ分析を行ってもらった．結果，両群の間に目的の有無が及ぼしたと考えられる差として，グラフの閲覧の仕方やグラフ閲覧中の発話に違いが現れた．

本研究の実験では，「既存商品・既存店舗についての弱みを知りたい」という分析目的を設定したが，この目的は「分析の観点」程度であった可能性がある．本研究の目的や意義の検証には分析で解決したい問題の定義や，問題解決のための分析目的をより詳細に設定する必要があり，今後の課題とする．

しかし，「分析の観点」程度でも参加者の思考や着眼点の特徴に違いが現れた．目的を与えられなかった目的なし群は，目の前にデータがあったとしてもそれに対する評価基準を持っていなかった．煩雑なデータを扱うほどに評価基準は必要であり，それを決めるタスクは非常に重くなる．

今後，実験環境を整理し，参加者を増やして再度実験を行えば，両群間の違いや「目的設定」の重要性をより明らかにできる可能性がある．

文献

[1] 藤本　玲子, 青山　幹雄, (2016)“データ駆動要求工学の提案とステークホルダ分析日の適用評価”, 研究報告ソフトウェア工学（SE）, Vol. 2016-SE-191, No. 15, pp. 1-8