

データマイニングとテキストマイニングの連携によるデータ分析支援

Data Analysis Support by Combining Data Mining and Text Mining

松本友哉 砂山渡 畑中裕司 小郷原一智

Tomoya Matsumoto Wataru Sunayama Yuji Hatanaka Kazunori Ogohara

滋賀県立大学 工学部

School of Engineering The University of Shiga Prefecture

Abstract: 近年では、アンケート分析などにデータマイニングやテキストマイニングの技術が用いられることが多くなってきた。しかし、データマイニングでは数値データ、テキストマイニングではテキストデータの分析を主とするため、その両者を含むアンケートデータ等の分析においては、用いるツールの選択が難しい状況があった。そこで本研究では、テキストマイニングのための統合環境 TETDM をベースとして、これにデータマイニングの機能を追加することで、両方向からのデータの絞込みとデータの解釈に役立てられる環境を提案する。

1 序論

近年では、アンケート分析などにデータマイニングやテキストマイニングの技術が用いられることが多くなってきた [1]。データマイニングは、大量の数値データから相関関係や法則性を発見することができるため、スーパーマーケットで一緒に購入されることが多い商品の組み合わせを調査などに利用されている。分析手法としては、頻出の組み合わせを発見するアソシエーション分析やデータ集合のまとまりを把握するためのクラスター分析などの手法が用いられている。一方、テキストマイニングは、大量のテキストデータから、文章の要約や否定的な意見の抽出ができるため、アンケート調査やコールセンターでの利用者からの意見の解析や分類に利用されている。分析手法としては、単語間のつながりを表す共起ネットワークや辞書登録をした言葉からポジティブな言葉とネガティブな言葉の判別する手法などが用いられている。そこで、データマイニングとテキストマイニングを組み合わせることにより、データマイニングでよく売れるルールを発見し、さらにテキストマイニングを使ってその消費者の意見を解析して理由まで発見するといった、両者の良い点を併せ持った分析の方法が考えられる。しかし、データマイニングでは数値データ、テキストマイニングではテキストデータの分析を主とするため、その両者を含むアンケートデータ等の分析においては、用いるツールの選択が難しい状況があった。

そこで本研究では、テキストマイニングのための統合環境統合環境 TETDM (Total Environment for Text

Data Mining) [2] をベースとして、これにデータマイニングの機能を追加することで、両方向からのデータの絞込みとデータの解釈に役立てられる環境を提案する。

2 関連研究

2.1 データ分析支援環境

データマイニングの分析支援環境としては、R[3] や Weka[4] がある。また、テキストマイニングの分析支援環境としては、「KH Coder」[5] や「UserLocal」[6] がある。これらのツールは、主に数値データまたは、テキストデータのどちらかのみを入力することを想定している。本研究では、数値データとテキストデータの2種類の入力に対応し、数値データとテキストデータ間で連携が可能な分析環境を開発する。

2.2 データマイニングとテキストマイニングの連携

「市場分析におけるテキストマイニングを活用したデータマイニングの実践」[7] という研究では、データマイニングでルールセットを発見し、テキストマイニングを用いてそれらのルールセットをフィルタリングする方法で連携がとられている。また、データマイニングから見たテキストマイニングの利用法に関する研究 [8] も存在する。これらの研究では、データマイニングまたは、テキストマイニングの片方が、もう片方を利用するという一方向の連携であるが、本研究では、

表 1: 本環境に使用する入力データの例

ID	年齢	性別	総合	外見	サイズ	テキスト
1	50	1	4	4	4	軽く、小さい、移動するのが便利です。
2	30	2	4	5	4	風量が弱い分音も静かです。
3	40	2	4	5	5	10 年ほど前に買ったヒーターの替えとして購入しました。
4	40	2	5	5	5	寒がりな私にはピッタリです。満足しています。
5	50	1	3	3	4	消し忘れに注意しましょう！コンパクトサイズでも暖房力あり！満足しています。
6	50	1	5	5	5	

データマイニングとテキストマイニングの双方向において絞り込み、分析ができる環境の開発を目指す。

3 データマイニングとテキストマイニングの連携分析環境

3.1 対象データとデータ形式

本環境では、数値データとテキストデータの両方を含むデータ集合を対象としている。例として、数値による選択回答と自由記述の両方を実施しているアンケートデータがある。本分析環境の入力におけるデータの形式の例を表 1 に示す。

3.2 分析環境の枠組み

本研究でのデータマイニングとテキストマイニングの連携の定義について述べる。本研究において連携とは、図 1 における、総合得点で 5 と回答したデータだけを抜き出すような数値情報による絞り込み、及び、テキストの中に特定単語が出現しているデータだけを抜き出すような単語によるデータの絞り込みを行い、絞り込んだデータの数値部分をデータマイニングツールに、テキスト部分をテキストマイニングツールへの入力とすることによる数値とテキストによる絞り込みと分析を組み合わせる分析方法を指す。

図 1 にデータマイニングとテキストマイニングの連携分析環境の枠組みを示す。

本環境は、テキストマイニングソフトの TETDM[2] に数値情報による絞り込み機能とデータマイニングツールを統計ソフト「R[3]」を用いて追加したものである。本環境における分析の流れを次に記載する。ま

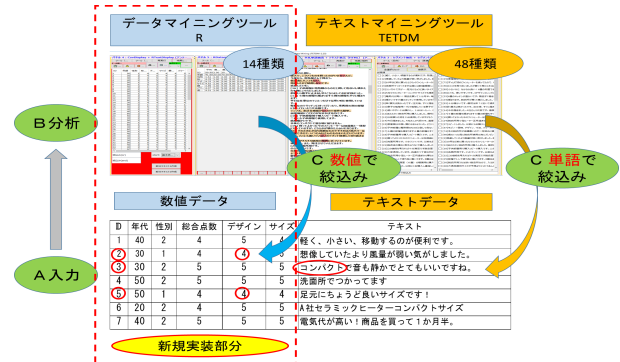


図 1: 本分析環境の枠組み

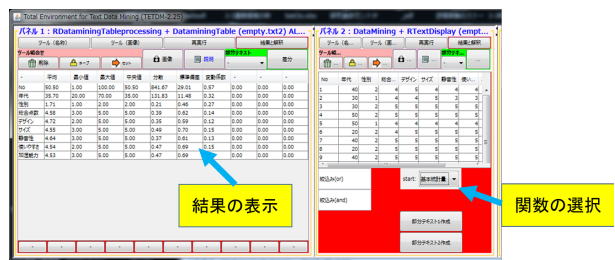


図 2: 「DataMining」及び「DataMiningTable」の操作画面

ず、図 1 の A にあたる部分でアンケートデータなどの数値とテキストがセットになったデータを入力する。図 1 の B にあたる部分では、A で入力したデータに対する分析結果が表示され、データマイニングツール 14 種、テキストマイニングツール 48 種を組み合わせる分析ができる。図 1 の C にあたる部分で必要に応じて数値、または単語でデータを絞り込むことができ、再度 A へ入力し、B にて分析を行うことができる。分析後に分析の結果から言えることを登録し、登録した知見を統合することで知見を獲得する。

3.3 入力データの形式

入力データは、属性と属性値のペアとして与えられるトランザクションデータとして与えられるものとする。そのうち、数値データ部分を抜き出した csv 形式のデータを数値データとして与え、テキスト部分はテキストファイルにまとめて入力として与える。

TETDM 上でのテキストの取り扱いは、以下のように分割して認識される。テキストデータ全体を「文章」、テキストデータ中の段落を「スナリバラフト」という文字列を段落間に挿入することで「セグメント」、句点で区切られた文を「文」として認識する。データを回答者などに区別する場合は、セグメントで分割する。

3.4 分析に用いられる機能

3.4.1 データマイニング用ツール

本研究では、TETDM と同じ Java に対応している点、一度作成した後も関数の追加が容易である点の2点から統計解析ソフトの「R」を用いてデータマイニング部分を TETDM に追加実装し、データマイニングとテキストマイニングの連携分析環境を構築を行った。そこで、データマイニング用ツールとして計算手法の選択に使用する先ほどの「DataMining」と計算結果を表示する「DataMiningTable」を作成した。「DataMining」及び、「DataMiningTable」の表示画面を図2に示す。

本ツールではデータマイニング機能として、基本統計量、基本統計量（フォーカス）、相関行列とアソシエーション分析 [9] を搭載している。基本統計量では、入力した数値データの列ごとの平均値、最小値、最大値、中央値、分散、標準偏差、変動係数を表示する。変動係数とは、単位の異なるデータ間のばらつきを変動係数同士で比較することで相対的に評価することができる数値である。基本統計量（フォーカス）では、入力した数値データの列全体の基本統計量に加え、着目しているデータのみを抜き出した基本統計量を追加して表示する。相関行列では、各列間の相関係数を行列形式で表示する。アソシエーション分析とは、例えば商店などで買い物が商品 A と商品 B を一緒に買う確率などを求めることができる分析である。

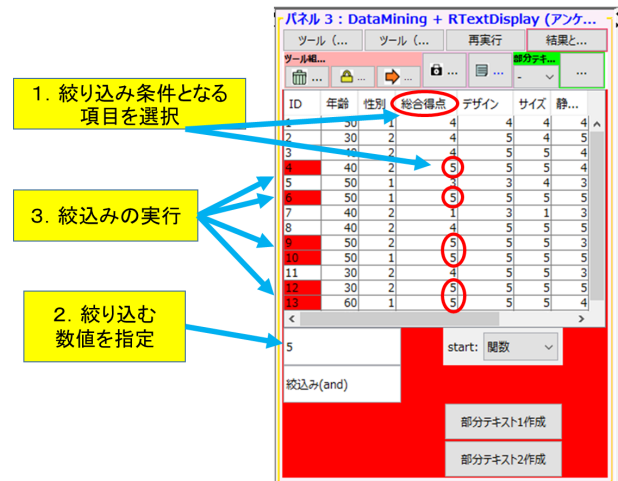


図 3: 数値情報によるデータの絞り込み用ツール

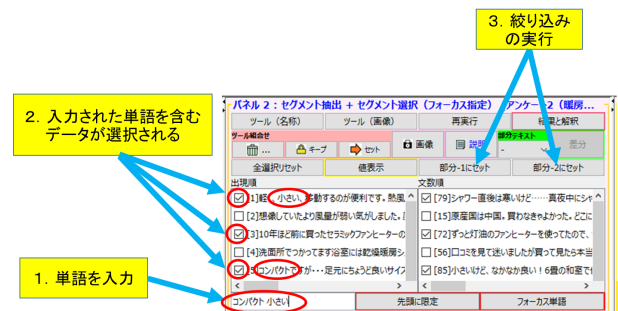


図 4: 単語情報によるデータの絞り込み用ツール

3.4.2 テキストマイニング用ツール

本分析環境の、テキストマイニングツールとして、TETDM の以下のツールを用いた。

- 「単語抽出」：指定した文字列をテキスト中でハイライト表示
- 「文章要約」：指定したキーワードに着目した文章の要約を表示
- 「失礼単語抽出」：否定的な表現をハイライト表示
- 「テキスト分類（再帰的クラスタリング）」：テキスト中の共起関係にある単語群を地図形式で表示
- 「単語情報まとめ」：単語の出現の頻度を表示

3.5 データの絞り込み機能

3.5.1 数値情報によるデータの絞り込み用ツール

数値データとテキストデータの両方を分析する上で、例えば、製品のどの部分が評価されているのかを分析するために、レビュー評価の高いテキストデータだけ

に絞り込みたい場合が想定される。そこで、TETDM 上で数値情報によるデータを絞り込むためのツールとして「DataMining」を作成した。「DataMining」の操作画面を図3に示す。

3.5.2 単語情報によるデータの絞り込み用ツール

数値データとテキストデータの両方を分析する上で、例えば、製品について「臭い」について言及している人の総合得点の平均などの数値情報や要約などを分析するために、「臭い」について言及している人だけにデータを絞り込みたい場合が想定される。そこで、TETDM 上で単語情報によるデータを絞り込むためのツールとして「セグメント抽出」がある。「セグメント抽出」の操作画面を図4に示す。

3.6 データマイニングとテキストマイニングの連携環境の構成

データマイニングとテキストマイニングの連携環境として用意した5つの組み合わせを以下で述べる。

3.6.1 「DM ↔ 原文参照」

「DataMining」の相関行列やアソシエーション分析により、レビューの総合得点と相関関係にある評価項目を発見し、「単語抽出」にてその評価項目で高評価や低評価を回答している回答者のテキストを絞り込んで原文を見ることでその理由を探ることが可能である。

3.6.2 「DM ↔ 要約」

「DataMining」の相関行列やアソシエーション分析により、レビューの総合得点と相関関係にある評価項目を発見し、「文章要約」によりその評価項目で高評価や低評価を回答している回答者のテキストを絞り込んで要約を見ることで絞り込んだデータの概要を知ることが可能である。

3.6.3 「DM ↔ 失礼単語着目」

「失礼単語地抽出」より、否定的な意見をハイライト表示させて否定的な意見や単語に着目し、「セグメント抽出」で着目した単語でデータを絞り込んで「DataMining」の「基本統計量」を見ることで、否定的な意見の男女比を見ることなどが可能となる。

3.6.4 「DM ↔ クラスタリング」

「テキスト分類（再帰的クラスタリング）」より、高評価につながる表現と共起関係のある単語に着目し、「セグメント抽出」で着目した単語でデータを絞り込んで「DataMining」の「基本統計量」を見てやることで、その単語に関係のあるの男女比を見ることなどが可能となる。

3.6.5 「DM ↔ 単語頻度」

「単語情報まとめ」より、出現頻度の高い単語など気になる単語に着目し、「セグメント抽出」で着目した単語でデータを絞り込んで「DataMining」の「基本統計量」を見てやることで、その単語に関係のあるの男女比を見ることなどが可能となる。

3.7 「DM ↔ 要約」の利用例

本節では、本環境内の「ツールセットのひとつである「DM ↔ 要約」の利用例を説明する。利用方法を説明する際の分析者の状況設定として、数値によるの評価項目と自由記述のある商品のレビューデータから、「ファンヒーター」に関するレビューの総合得点が高い回答者は、どの評価項目を重視しているのか、また、その理由を知りたい場合を想定する。以下に想定した場合の目的を達成するための「DM ↔ 要約」の具体的な分析例を示す。そのときのパネルの様子を図5に示す。

1. レビューデータの数値データ部分に該当する csv ファイルを TETDM フォルダの「csvfile」フォルダ内に格納する。
2. レビューデータのテキストデータに該当する txt ファイルを TETDM フォルダの「text」フォルダ内に格納し、メニューウィンドウ左上「テキスト入力」、「ファイル」ボタンより選択する。
3. ツール「RApplication」の「DM ↔ 要約」ボタンを押して「DM ↔ 要約」をセットする。
4. 「DataMining」のアソシエーション分析で、総合得点が5と一緒にでてくるルールセットを探す。
5. アソシエーション分析をした結果、全体の46パーセントの人が「総合得点」の評価が5点であると同時に「使いやすさ」の評価が5点で回答していることがわかる。
6. 「DataMining」の数値情報による絞り込み機能を用いて、「使いやすさ」の評価を5と回答した回答者のみデータを絞り込む。
7. 「部分テキスト作成1」ボタンを用いて、「使いやすさ」の評価を5と回答したデータのみのテキストデータに加工する。
8. 「文章要約」パネル右上の部分テキストから、「部分テキスト1」をセットし、先ほど絞り込んだテキストデータを「文章要約」の入力データとする。
9. 「文章要約」のキーワードの中から「使いやすさ」という観点から連想される「便利」、「機能」といったキーワードに着目して要約を行う。
10. 要約結果から、「余計な機能が無くわかりやすい」点が評価されていると推測できる。

これらの手順から、例えば、商品の購入者は「使いやすさ」を重視しており、この商品は、余分な機能が無くわかりやすいことが高評価につながっていると解釈できる。

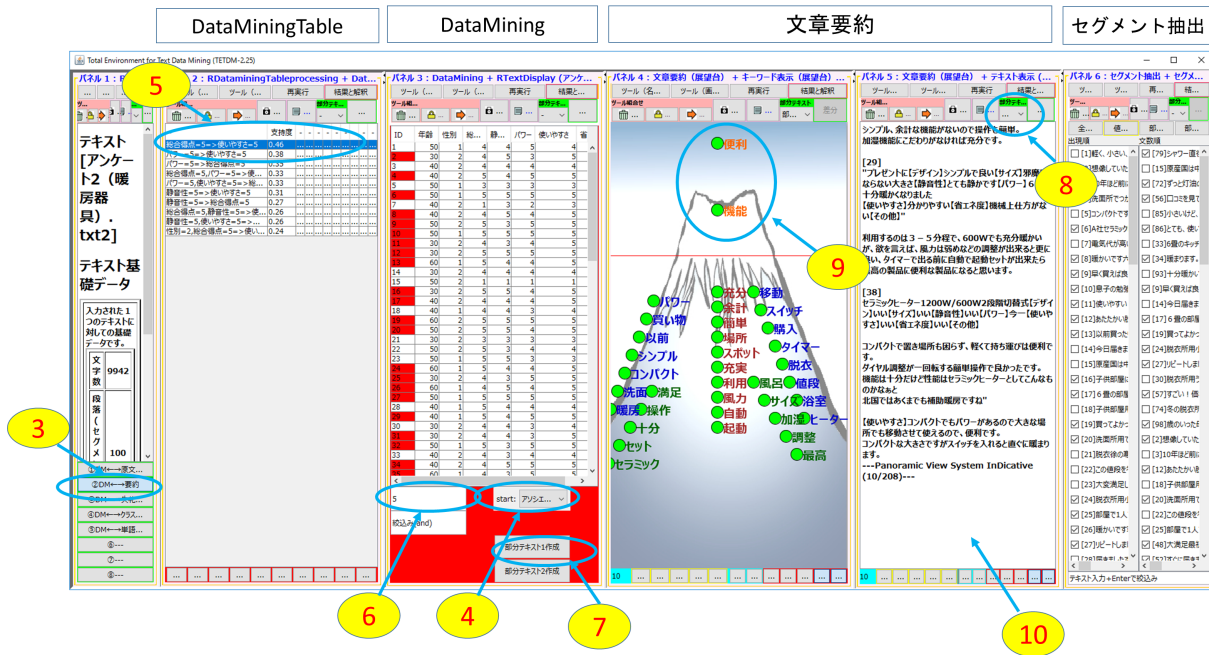


図 5: 「DM ↔ 要約」の利用例

4 評価実験

本章では、本環境の使用によるアイデアの具体性の変化を検証した実験について述べる。

4.1 実験内容

本節では、分析環境の評価実験内容について述べる。提案分析環境によって得られる知見の着目点の有無と具体性の有無を評価するために実験を行った。被験者は、大学生・大学院生の計 16 名を対象に、後述する指定の分析環境を用いて、商品のレビューデータを分析し、「利用者の満足度（レビューの総合点数）が高くなると予想される、新製品を提案」をしてもらった。提案環境で分析を行う A グループと、比較環境で分析を行う B グループの 2 グループを 8 名ずつに分けて実験を実施した。提案環境の A グループは、本研究の連携機能がある単語、数値で絞り込んだデータをデータマイニング、テキストマイニングの両方で分析できる環境で 3.6 で紹介したツールセット 5 つを使用できる。一方で比較環境の B グループは、本研究におけるデータマイニングとテキストマイニングの連携機能が無い、3.6 で紹介したツールセットをばらしたデータマイニングツール 1 種、テキストマイニングツール 5 種の計 6 つのツールを連携機能が無く個別に実施できる環境下である。今回の実験で使用するアンケートデータは、商品の機能やデザインなどを評価する 5 点満点の数値とレビューテキストがセットになった形式とし、「楽天市

場」[10] で販売されている「加湿器」と「ファンヒーター」のレビューデータを各 100 件を使用した。被験者には、商品のレビューデータより、「利用者の満足度（レビューの総合点数）が高くなると予想される、新製品の提案」を目的として、以下の手順で行ってもらった。

1. データを数値や単語によって絞り込む。
2. 分析に用いるツールをセットする。
3. 分析して発見した「結果と解釈」を登録する。
4. 手順 1 から 3 を繰り返した上で、集めた「結果と解釈」をまとめ、新製品の提案を行う。

4.2 最終提案における着目点の有無と具体性の有無の評価

本節では、最終提案における着目点の有無と具体性の有無の評価の方法と考察について述べる。データマイニングとテキストマイニングの連携の有無により、最終提案における着目点の有無と具体性の有無の評価を行った。着目点の有無とは、アイデアが性別や年齢といったどの顧客層に対して向けられたものなのか判断できる場合は着目点有とした。また、改善に関する複数の概念のうち他よりも重要度が高い旨の記述がある場合も着目点有とした。具体性の有無の評価については、例として単に「デザインを良くする」と言った提案だけでなく、デザインの方向性や程度が記載されて

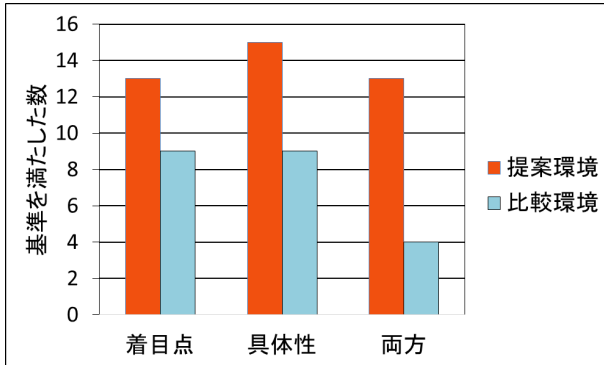


図 6: 最終提案における着目点の有無と具体性の有無の評価の集計結果

いる場合は、具体性有と評価した。提案環境と比較環境における「加湿器」と「ファンヒーター」の分析結果を評価し、集計したものを図 6 に示す。

図 6 より、提案環境の方が着目点有の数、具体性有の数、両方有の数の全てにおいて上回っている。提案環境のアイデアの一例を挙げると「購入者は女性が多く、また、女性はデザインを重視していたため、デザインを可愛らしく凝ったものにするべきである。」といった提案があった。「女性が多く」、「デザインを重視していた」という点からアプローチする客層や他の評価項目よりも重視していることがわかるため、着目点有と評価した。また、「デザインを可愛らしく凝ったもの」の点から、デザインの方向性や程度がわかることから、具体性有と評価した。これより、データマイニングツールとテキストマイニングツールを個別に使用するよりも、連携して得られた知見のほうが、着目する箇所が示され具体的な内容になったことが確認できる。

次に、提案環境と比較環境におけるデータマイニングツールとテキストマイニングツールの使用頻度の偏りと最終提案の評価点の平均を比較し、表にまとめたものを表 2 に示す。ただし、ツール使用頻度の偏りは、TM の利用率と DM の利用率との差を表し、最終提案点数は、被験者の各提案に対して着目点有で 1 点、具体性有で各 1 点と評価し、「加湿器」と「ファンヒーター」の提案の評価点を合計したものを表す。

表 2 より、提案環境のほうが、ツールの使用頻度の偏りが比較環境より 0.37 低く、最終提案の評価点の平均が比較環境より、1.2 点高いことがわかる。以上より、提案環境のほうがデータマイニングとテキストマイニングツールをバランスよく使用できるため、着目点が明確になり、かつ具体的な知見を得ることができた。

表 2: 提案環境と比較環境のツール使用頻度の偏り及び分散と最終提案の評価点の平均

	ツール使用 頻度の偏り の平均	ツール使 用頻度の 分散	最終提案 の評価点 の平均
提案環境	0.42	0.6	3.5
比較環境	0.79	0.19	2.3

5 結論

本研究では、数値データ、テキストデータの両方を含む入力に対し、データマイニングとテキストマイニングの両方向からのデータの絞り込みと分析が行える環境を構築した。本環境で分析することで、得られる知見の具体性が向上することを実験により検証した。

参考文献

- [1] 佐藤弥生, 佐々木千晶: 介護職員の「不適切ケア」の判断の拠り所-アンケートの自由記述の分析から-, 岩手県立大学社会福祉学部紀要第 18 巻, pp.11 – 22, (2016)
- [2] Total Environment for Text Data Mining : (URL)<http://tetdm.jp>
- [3] GNU R : (URL)<https://www.r-project.org/>
- [4] Weka (URL)<http://www.cs.waikato.ac.nz/ml/weka/downloading.html>
- [5] KH Coder : (URL)<http://khc.sourceforge.net/>
- [6] UserLocal : (URL)<http://textmining.userlocal.jp>
- [7] 竹内広宜, 杉山喜昭, 山口高平: 市場分析におけるテキストマイニングを活用したデータマイニングの実践-生ごみ処理機の市場分析を例として-, 知能と情報, Vol.24, No.3, pp.728 – 742, (2012)
- [8] 善田昌樹: データマイニングの視点から見たテキストマイニングの三つの利用法, 日本情報系学会誌, Vol.35, No.1, pp.4 – 18, (2014)
- [9] 荒木孝治: R と R コマンドで始める多変量解析, p.180, 日科技連, (2007)
- [10] 楽天市場 : (URL)<http://www.rakuten.co.jp/>