公立はこだて未来大学 2014 年度 システム情報科学実習
グループ報告書

Future University-Hakodate 2014 System Information Science Practice

Group Report

プロジェクト名

データ解析技術による意思決定支援

Project Name

Support for decision making by data analysis

グループ名

グループ B

Group Name

Group (B)

プロジェクト番号/Project No.

18-B

プロジェクトリーダ/Project Leader

1012154　杉澤智己　Tomoki Sugisawa

グループリーダ/Group Leader

1012106　井川翼　Tsubasa Ikawa

グループメンバ/Group Member

1011213　岩橋賢吾　Kengo Iwahashi
1012047　田中桂介　Keisuke Tanaka
1012062　山田林太郎　Rintaro Yamada
1012106　井川翼　Tsubasa Ikawa
1012131　祐川翔斗　Shoto Sukekawa
1012135　田中健介　Kensuke Tanaka
1012156　中島大貴　Hiroki Nakashima
1012173　小林美沙　Misa Kobayashi

指導教員

片桐恭弘　竹之内高志　永野清仁

Advisor

Yasuhiro Katagiri　Takashi Takenouchi　Kiyohito Nagano

提出日

2014 年 1 月 14 日

Date of Submission

January 14, 2015

概要

本プロジェクトのグループ B の目的は、機械学習技術を用いて本学の 2 年進級時におけるコース選択の意思決定支援システムを構築、製作することである。はこだて未来大学では、2 年に進級する際に 4 つの中からコースを選択して進級する必要がある。そのコースの選択における意思決定は人によっては困難なものであり、どのコースにするべきかと迷うことも少なくない。そこで我々グループ B はコース選択における意思決定支援システムを製作することで、本学の 1 年生の進級時によりよいコースを提示し、コースの決定を円滑、かつ的確にすることを目指す。本プロジェクトでは、本学の生徒の傾向を様々な方法で調べそれをデータ化し、その大量のデータを機械学習技術を用いて解析することにより理論的なコース選択支援システムの構築を目標とする。これらの実現のために我々はデータ収集班とデータ解析班の二つにわかれ協力してこれを実現することとした。まず、データ解析に使う予定である R 言語の知識習得から行った。MeCab という自然言語解析ソフトを有効に活用することにより、技術習得のための練習用として多くのニュース記事を内容からジャンル分けすることからはじめ、文書解析技術を学んだ。それと平行し、コース選択の意思決定支援システムに必要である入力データや出力データの検討、ユーザインターフェースの構想やデータ収集、解析の方法など様々な事柄について議論、検討を十分に深めた。データ収集班の目標としては、必要な入力データの検討と収集、データ解析班の目標としては解析技術の調査、検討ということが決定した。各班の議論の結果、4 つのコースの 2 年生から 4 年生に向けてアンケート調査を行い、その回答から機械学習により支援システムを完成させるということとなった。また、この支援システムは「どのコースに適しているか」という判別の役割しか担うことができないため、コース選択の一助となるような情報を載せた HTML のページを作り、システムの補助として判別の結果と共に表示するという形式でプロジェクトを進めることとした。また、支援システムには機械学習技術だけでなく、因子分析などのデータ解析技術も同時に使用して精度を高めることとなった。

キーワード　文章解析, 機械学習, R 言語, 意思決定支援, 因子分析

（文責: 井川翼）

Abstract

Abstract in English. The purpose is to develop course recommendation system with machine learning for 1st grade students in our University, who should select a course. The selection can be difficult or agonizing for some people. Then, we set out the development of course recommendation system to get the selection easier and more correct. Our goal in this project is to develop a logical course recommendation system with analysis of machine leaning of trends students in our university found by some kind of means. We organized two groups, Data gathering team and Data analysis team, for our goal. First, we studied R language, which would be utilized for data analysis. We learned method of text analysis through categorizing news articles with Mecab using a natural language analysis software as practice. At the same time, we discussed followings for development the system. ・Input and output data ・Framework of user interface ・A means of data gathering or analysis We decided that Data gathering team consider what king of data we need and get the dataset, and Data analysis group search and investigate the analysis methods. As a result of the discussion, we decided to have questionnaire for students from 2nd to 4th grade for the system development. In addition, we were gong to make HTML that show the information of courses to compensate the system ability giving advice about which course is suitable. Furthermore, we decide to take another data analysis method like factor analysis but not only machine learning to elaborate.

Keyword data analysis, machine learning, R (programing language), support for making decision, Factor analysis

（文責: 岩橋賢吾）

目次

第 1 章　背景

本学は入学して 1 年後、2 年生に進級する際に進むコースを選択しなければならない。コースによって様々な特色があり、ガイダンス等でそれらの説明はなされる。しかし、往々にして「進むべきコースに迷う」という学生も少なくないのが現状である。シラバスに講義内容が掲載されているが、それを見ないで決めてしまう学生も居る。やりたいことの有無や、成績状況、様々な要因から、必要な情報のみを選び出してよりよいコースを選択する。その選択は我々がそうであったように、人によっては困難なものであると予想できる。

（文責: 井川翼）

1.1　該当分野の現状と従来例

1 年生の進級時、コース選択における様々な情報の取捨選択が難しく、コース選択がスムーズにできない。また、情報を得る場所が多岐に渡り、何もしないでいると得られない情報も出てくる。自分に見合ったコースがどこであるのかということがわからない人がいる。コース選択のために 1 年生には十分な選択期間が設けられ、ガイダンス等もあるが、それらを生かしきれずにコース選択に悩む学生もいる。

（文責: 井川翼）

1.2　課題の概要

本学の生徒からデータを収集し、それらのデータを解析することによりコース選択における意思決定を支援してくれるようなシステムを製作する。データの収集方法や解析方法について詳しく議論、検討を行い的確で理論的なシステムの完成を目指す。

（文責: 井川翼）

第 2 章　到達目標

2.1　本プロジェクトにおける目的

本プロジェクトの目的は、コース選択における意思決定支援システムを構築することである。本学の 1 年生がコースを選択するために得るべき情報を 1 元化し、その学生の情報からさらに最適なコースを選択、表示してくれるような機能を持ったシステムを構築していく。システム UI についても、入力や出力に面白みを持たせることで、積極的な利用を促し、自然な使用ができるような工夫も施していく。最終的にはそれらの UI や機械学習とデータマイニングを利用したシステムを組み合わせ、コース選択支援システムを完成させる。

（文責: 岩橋賢吾）

2.1.1　通常の授業ではなく、プロジェクト学習で行う利点

通常の授業ではどうしても個人作業となる。プロジェクト学習においては、多人数で作業できることが最大の利点となる。本プロジェクトでは、システムの構成にはどうしても多くの人の意見を取り入れていく必要性がある。また、データの収集や解析など、どうしても一人二人では解決できない量の作業がある。それらの作業を分担し、個々ではなくチームとして作業することで効率よくシステム作成が行える。また、システムの改善点を見つけるなどの作業も、プロジェクトのような多人数であれば多くの意見を元に改善をすることができる。

（文責: 岩橋賢吾）

2.2　具体的な手順・課題設定

コース選択における意思決定支援システムを完成させるために、我々は初めに技術習得を目標とした。その技術習得の後に、本格的にコース選択支援システムを製作していく。以下のように手順を設定した。

1. R 言語の基礎的な学習
   課題：R 言語とはなにか。そして、R 言語を使用してどのようなデータを処理し、どのような結果が得られるのかを全員で把握、習得する。
2. テーマ議論
   課題：新規の本プロジェクトにおいて、どのような目的を設定するのかを議論する。全員でテーマ案を持ち寄り、それらについてブレインストーミングを行う。
3. 新聞記事の収集
   課題：データ収集、解析技術の習得のための練習課題として、前期までに新聞記事の分類システムを製作する。その製作にあたって必要な新聞記事の収集を行う。
4. 新聞記事の解析
   課題：データ収集、解析技術の習得のための練習課題として、前期までに新聞記事の分類シ

ステムを製作する。その製作にあたって収集されたデータを解析する。

5. コース選択支援システムの具体的な検討
   課題：これまでに習得した技術などを元に、さらに適当なシステム製作についての話し合いを行う。システムの入力、処理、出力や、妥当性などを検討する。
6. 学生へのアンケート
   課題：本プロジェクトのシステムを完成させるために必要不可欠な学生へのアンケートを行う。対象者や、とるべきアンケートの内容についても具体的に議論する。
7. システムに使用する技術
   課題：システムを製作するに当たって、様々な既存の技術の中から適当なものを選んで使用する必要がある。そのため、使用する技術について議論、検討する。
8. 収集したデータの処理方法の検討
   課題：収集された学生へのアンケートを、複数の機械学習手法を使って処理する。それらの結果から、より良い機械学習の手法を検討する。
9. システムの試作、テスト
   課題：集まったデータよりシステムを試作する。試作されたシステムを被験者を集めてテストし、フィードバックをもらう。
10. システムの改善
    課題：得られたフィードバックよりシステムの改善案を議論、検討してシステムを改善する。

11. 完成したシステムのテスト
    課題：完成したシステムを実際に使用してもらい、使いやすさなどをフィードバックしてもらう。

（文責: 小林美沙）

第 3 章　課題解決のプロセス

3.1　基礎知識と技術の習得

本プロジェクトでは、機械学習技術を応用したシステムの開発を目指す。この機械学習技術は統計学との関連が深いため、その統計的なデータ解析にしばしば用いられる R 言語を使用する。R 言語には統計解析の関数群をはじめとした非常に多くの機械学習アルゴリズムが実装されている。課題解決に必要な機械学習がどいういうものであるかを知るため、R 言語の学習をグループメンバー全員で行った。

（文責: 田中健介）

3.1.1　天気の予測

練習用として、1 年間の天気データを使用して練習を行った。天気データは以下のような形式の csv ファイルにまとめられている。このデータを R 言語環境で読み込み、データ処理を行う。

日付	気温	風速	気圧	湿度	天気
2013/5/17	11.8	3.1	9.4	71	晴
2013/5/18	12.5	2.7	9.7	68	薄曇
2013/5/19	11.2	2.3	9.5	72	曇
2013/5/20	10.9	1.9	12	92	雨時々曇
…	…	…	…	…	…
2014/5/17	11.4	7.4	8.1	60	曇

表 3.1　天気サンプル

ここでの目的は、「気温、風速、気圧、湿度」よりその日の天気を予測するようなシステムを作ることである。データ処理の手順を次に示す。

1. データから、「日付」データを消去する。
2. 欠損したデータは、扱いを簡潔にするため取り除く。
3. 天気の情報を数値として置き換え、機械学習で処理できるようにする。
4. ここまででつくられたデータより、機械学習を行う。
5. 機械学習によって作られた関数で天気を予測する。

手順の 4 つ目の機械学習には、様々な手法がある。今回は、LDA(線形判別分析) と SVM(サポートベクターマシン) の 2 つの手法を用いる。

（文責: 田中健介）

Latent Dirichlet Allocation(線形判別分析)

LDA とは、ある複数のグループに分けられたデータをなるべく誤判別の少ないように線形的に分類する手法である。

実際に R 言語を用いて LDA を使用し天気の予測を行ってみると、おおよそ 80% の予測に成功した。

（文責: 井川翼）

Support Vector Machine(サポートベクターマシン)

SVM とは、教師あり学習を用いた識別手法のひとつである。これは現在知られている多くの手法の中で一番認識性能が優れた学習モデルの一つである

実際に R 言語を用いて SVM を使用し天気の予測を行ってみると、おおよそ 83% の予測に成功した。

（文責: 井川翼）

3.1.2 文章解析

本プロジェクトでは R 言語を用いた機械学習技術を応用して、意思決定支援を行うことを前提に機械学習技術の習得を行っていた。我々データ解析班は加えて、文章を解析するための技術を学習した。文章解析、つまり自然言語処理のためには形態素解析を行う必要がある。そのため、R 言語で使用できる形態素解析ソフト「RMeCab」を学習することとした。学習には「R によるテキストマイニング入門.」[1] を使用。主だった形態素解析、そしてテキストマイニング技術を習得した。

（文責: 井川翼）

3.2 課題決定のプロセス

R 言語の機械学習やテキストマイニングなどの技術習得の後、本格的にテーマ案の議論を開始した。テーマ議論は

1. テーマに関する方向性について再確認する (機械学習、テキストマイニング、意思決定支援)。
2. 各々がテーマ案を提案する。
3. テーマ案について、それぞれ入力や出力、利点や欠点、対象ユーザや実現性などについて話し合う。
4. 話し合ったテーマ案について改善案や、代替案などを出すことで具体化を進める。
5. 最終的にテーマを決定する。

という流れで行った。議論の中で最終的に「SNS を利用したもの」や「オススメのコンテンツ」、「コース選択支援」が残った。この中で、新規性がありかつ本プロジェクトの内容に沿ったものということで「コース選択支援」をテーマとすることが決定した。

（文責: 田中健介）

3.3 課題解決のための練習

コース選択支援をテーマとしてプロジェクトを進めることが決定したが、具体的にどのようにして実現すればよいのかを知って活動に臨むため、まずは最終目標の練習としてニュース記事の分類という練習課題を設定した。具体案は以下の通りである。

1. ニュース記事を Web サイトから収集する。
2. 収集するデータは形式を一律に揃え、データの処理を行いやすいようにする。
3. データ解析はひとつの方法に限らず、複数の方法を模索していく。
4. 分類器を作成する。

（文責: 小林美沙）

3.3.1 新聞記事の分類

データの収集

新聞記事については、Web サイト [2] にて以下のようなものを収集した。また、保存形式は txt で、タイトルを「ジャンル名-0000.txt」というような形式にフォーマットを揃えることで、データ処理を行いやすいように配慮した。記事数は全部で 1725 である。

ジャンル	ジャンル名	記事数
政経	eco	275
社説	edi	52
選挙	ele	90
環境	env	22
五輪	fiv	140
IT	inf	102
政治	pol	454
地域	reg	73
社会	soc	87
スポーツ	spo	298
科学	tec	59
国際	uni	110

表 3.2　収集した新聞記事

これらを次の手法でデータ解析していく。

（文責: 小林美沙）

最大エントロピー法

最大エントロピー法は、分類を行うときに使われる手法の一種である。分類されるときの確率を求めるとき、与えられた制約の中でエントロピーを最大化するような手法であり、これにより未知

のデータに対して確率をなるべく一様に分配することができる。これは「ゼロ頻度問題」という、未知のデータが出てきた瞬間に確率が 0 になってしまうような問題に強い手法である。この手法を使って次のような手順でデータを処理した。

1. ニュース記事を、RMeCab を使って形態素解析を行い、「名詞」「動詞」情報を取り出す。
2. これらの単語情報を、1 記事ごとにまとめる。
3. 全ての記事のうち 7 割を学習データ、3 割を予測データとしてランダムに分配する。
4. 最大エントロピー法により、分類を行う。

以下に、分類結果を示す。

		正しい分類											
		政経	社説	選挙	環境	五輪	IT	政治	地域	社会	スポーツ	科学	国際
分類結果	政経	61	3	1	0	2	5	11	5	4	2	2	2
	社説	4	10	0	1	0	0	3	1	0	0	0	0
	選挙	0	0	4	0	1	0	16	0	0	0	2	0
	環境	1	0	0	2	0	0	0	2	1	0	3	0
	五輪	4	0	0	0	25	0	2	0	1	0	0	0
	IT	4	1	0	0	0	25	0	0	0	1	0	0
	政治	3	1	20	1	5	1	83	6	4	7	1	5
	地域	1	0	0	0	1	0	0	1	1	0	0	0
	社会	6	0	1	0	1	0	2	5	8	0	1	0
	スポーツ	6	0	0	0	2	0	3	3	1	79	0	1
	科学	1	0	0	0	0	1	0	2	2	1	7	0
	国際	1	3	0	0	1	0	2	0	1	0	1	22

表 3.3　最大エントロピーによる分類結果

この結果から、全体の正答率はおよそ 64% であるということがわかる。理論的に、ランダムで分類したときの正答率はおよそ 8% 程度であることを考えると、正しい分類ができているといえる。

（文責: 井川翼）

LDA(線形判別分析)

LDA は分類を行うときに使われる手法の一種である。特徴の線形結合の値に基づいて分類を行う確率的分類器で、グループ分けの境界が直線、あるいは超直面であり、線形関数を用いてグループの所属の判別を行う手法である。この手法を使って次のような手順でデータを処理した。

1. ニュース記事を、RMeCab を使って形態素解析を行い、「名詞」「動詞」情報を取り出す。
2. これらの単語情報を、1 記事ごとにまとめる。
3. 全ての記事のうち 7 割を学習データ、3 割を予測データとしてランダムに分配する。
4. LDA により、分類を行う。

以下に、分類結果を示す。

		正しい分類											
分類結果		政経	社説	選挙	環境	五輪	IT	政治	地域	社会	スポーツ	科学	国際
	政経	53	0	0	2	0	6	4	3	2	1	1	1
	社説	1	9	0	2	0	1	0	0	0	0	2	0
	選挙	1	0	16	0	0	0	14	0	0	0	0	0
	環境	0	0	0	3	0	0	0	4	1	0	0	0
	五輪	1	1	0	0	40	0	2	0	1	0	1	0
	IT	1	1	0	0	0	20	1	1	0	2	2	0
	政治	4	0	23	0	2	3	84	2	1	0	0	3
	地域	4	0	0	0	0	0	2	11	6	1	0	0
	社会	5	0	0	0	0	2	2	11	6	1	0	0
	スポーツ	2	0	0	0	0	1	0	1	0	90	0	0
	科学	3	0	0	0	0	0	2	1	0	1	10	0
	国際	1	0	0	0	0	4	1	0	2	0	0	18

表 3.4　LDA による分類結果

この結果から、全体の正答率はおよそ 71% であった。こちらも最大エントロピー法と同様に、正しく分類できているといえる。

（文責: 井川翼）

考察

この 2 つの結果から、2 つの分類器はある程度の精度をもって機能していることがわかる。また、誤分類された例の内訳を詳しく見てみると、「選挙」についての記事が「政治」についての記事として誤って分類されているものが多かった。これは内容が似通っていて、人間でも分類に迷うような内容であることが原因なのではないかと思われる。また、データ数が極端に少ないジャンルの記事（たとえば、「環境」などは 1725 の記事のうち 22 ほどしかなかった。その中から 7 割を学習データとしたので、おおよそ 15 前後しか学習データがなかった）についてはやはり精度がよくなかった。逆にデータ数の多いものは、よい精度で分類される傾向がみられた。

大まかな分類としては成功したが、似通ったデータ、数の少ないデータに対する工夫をすることでもっと精度が上げられたと思われる。

（文責: 井川翼）

3.3.2　コース選択支援システムデモ

コース選択における意思決定支援システムのデモを作成するため、本学 3 年生 79 人にアンケートを行った。内訳は以下のとおりである。

また、アンケートの内容を以下に示す。

0. 所属コース

1. 好きだった科目

コース	人数
情報デザイン	20
情報システム	18
知能システム	22
複雑系	19

表 3.5　アンケートの内訳

2. 入学方法
3. 必修科目の A の数
4. 落とした必修科目
5. 将来の夢
6. 第一志望だったコース
7. 習得済み単位数
8. 卒業後は
9. 最履修の科目数
10. 講義中に携帯電話または PC で SNS を使用したことがある
11. 講義中に寝てしまうことがある
12. 講義のノートはしっかりとる
13. 試験は一夜漬けではなくしっかり計画して勉強する
14. プロジェクトリーダーまたはサークルの部長をやっている
15. 学校に泊まることがある
16. 友達は多いほうだ
17. 大学が楽しい
18. 未来大学に来てよかった
19. プログラミングが好き
20. 車を持っている
21. 出席を取らない講義にでもきちんと出席している
22. 絵を描くのがすき
23. おしゃれが好き
24. 寝坊が多い
25. 現在恋人がいる
26. ギャンブルがすき
27. 眼鏡をかけている
28. タバコをすう
29. お酒が好き
30. 運動がすき
31. 血液型
32. 現在アルバイトをしている
33. 友達と遊ぶ回数
34. 勉強は 1 人でする
35. 各コースの印象

36. 髪を染めていますか
37. 推奨気ではない PC を使用している
38. 工作が好き
39. 所属コースへの満足度
40. わからない問題を先生のところに聞きに行く
41. 課題をしっかりやっている
42. サークル、部活
43. ピアスが開いている
44. 好きな色
45. アニメが好き
46. ゲームが好き
47. 住んでいる場所
48. 所属コースで必要なスキル、勉強
49. ボランティアに参加したことがある

はじめに、この質問の中で、処理のしやすい Yes or No で回答できるものを抽出。その後、ランダムフォレストという機械学習手法で分類を行い、コースを結論付けるために重要である質問を抽出する。質問を 10 であるとして、機械学習にかける 10 の質問の組み合わせを順に変更していき、分類精度の上がる質問をピックアップしていく方法を取った。

その結果

15. 学校に泊まることがある
16. 友達は多いほうだ
19. プログラミングが好き
20. 車を持っている
21. 出席を取らない講義にでもきちんと出席している
22. 絵を描くのがすき
28. タバコをすう
32. 現在アルバイトをしている
34. 勉強は 1 人でする
38. 工作が好き

という 10 の質問に絞ることができた。これらを使って中間発表時にアンケートを行い分類器で予測した結果、その人が実際に所属しているコース、あるいは所属したいコースであるかを確かめた。正答率は 4 割前後と、ランダムに判断するよりはよい程度の結果となった。

（文責: 井川翼）

3.4 課題解決過程の詳細

3.4.1 システムの検討

中間発表にてデモとして公開したコース選択支援システムに寄せられた意見を元に、どのようにしてシステムを構築していくかを検討した。

分類システムに使用するデータ

まず、コースを分類するにあたってどのようなデータを使うべきかを話し合った。これは、何を根拠とするかで分類結果が変わる、システム作成における重要なポイントである。前期の最後から話し合いをしていた議題であり、案としては、学びたいことを入力としそれが学べるようなコースをお勧めする、なりたい職業や就職先を入力としそれが実現できるようなコースをお勧めする、などが挙げられていた。後期ではさらに具体的に話し合いをし、2〜4 年生の実態を調査し、その特徴に当てはまる人を各コースに分類するというものが挙げられた。しかし、これで本当に適切なコースに分類できるのか、各コースに多く見られる特徴に当てはめているだけではないかという疑問が生じた。そこで、各コースの特徴を知るために学生にアンケートを実施する他、各コースの教員にもアンケートを行うことを検討した。学生にとるアンケートの内容の案として、一週間の勉強時間、好きな言語、どの科目をどのくらい好きでその成績はどうだったか、借りた本の冊数、履修した選択講義数、認知科学で書いたレポートのテーマ、身についた能力、インターンで役立ったこと、卒研テーマ、就活ではなしのネタになったこと、コースで学べたこと、志望している職種、有意義だった講義、入学動機、これからやりたいこと、が挙げられた。教員にとるアンケートの内容の案として、やる意義が感じられなかった講義、そのコースに必要な能力、が挙げられた。

アンケートについて

次に、アンケートの内容とデータ数について話し合った。このシステムを利用する 1 年生は、どのコースに行くべきか迷っている人、自分が何に向いているのかがわからない人を想定している。学びたいことが既に的確である人、入学動機や将来の目標がはっきりしている人はこのシステムを利用しないと考えたためである。そのため、どのようなことが学びたいか、どのような目的で入学したのか、などを入力としてコースを分類するのは、自分のやりたいことがわかっている人にとっては単なる再確認であり、迷っている人やわからない人にとっては利用しにくいシステムになってしまうと考えた。具体的にやりたいこと、というような入力はできるだけ避け、興味や関心があるかないかを 5 段階で問うようにした。学生にとるアンケートの内容については、コースで学べたこと、身についた能力を問うような質問をする方針に決定した。本プロジェクトでは、機械学習を用いてのコース分類器の作成が目標であるので、そのために同じアンケートをシステム利用の際に 1 年生にも回答してもらうことになる。そのため、身についたことではなく興味関心を問うような質問に変更した。具体的なアンケート内容に関しては、シラバスから各コースの講義内容の頻出キーワード分析をし、暫定的な上位 30 単語ずつを調査してアンケート作成の参考にした。

サンプルサイズ

どのくらいのデータがあれば適切な分類器の作成ができるかを話し合った。まず、各学年の人数を 240 と仮定し、 1 年生には過半数の 120 人にアンケート調査を行うことを目標とした。2〜4 年生には、情報システムコース 95 人、デザインコース 45 人、知能システムコース 70 人、複雑系コース 70 人の計 280 人にアンケート調査を行うことを目標とした。前期に行った新聞記事の分類での訓練データとテストデータの比率が 7:3 であり、テストデータとなる 1 年生のデータを 120 としたとき、訓練データは 280 必要となることがわかる。はこだて未来大学公式ホームページによると、各コースの人数は、情報システムコース、デザインコース、複雑系コース、知能コースの順に、80 人,40 人,60 人,60 人である。これより、280 人をこの比率に割り振ってでてきた数が 95 人,45 人,70 人,70 人である。

システムの補助について

最後に、システム以外のことについても検討した。コース診断だけではなく、他にも提供できる情報はないかを話し合った。これは、おすすめのコースを知るだけではなく、興味のありそうな講義や卒業研究のテーマの情報を提供し、各コースのことをさらに理解してもらえるようにするためである。コースについて何もわからなかった時に、どのような情報があったらよかったかを考え案を出し合った。各コースの簡単な紹介はもちろん、各コースの教員の情報や、必修科目・選択科目の講義内容の紹介、プロジェクト学習の紹介、卒業研究の紹介、先輩方の就職先の情報が挙げられた。他にもあるのではないかと考え、1 年生に各コースについて知りたい情報を調査し、その情報を掲載することにした。

（文責: 小林美沙）

3.4.2 アンケートの作成

作成したシステムにおいて機械学習を行うために必要なデータとしてアンケート収集を行った。そのためのアンケート作成について述べる。

初期構想

アンケートを実施するにあたって初期段階ではどのような設問であればコースごとの特色が現れるのかを見極めるための予備調査アンケートを実施、その後に機械学習にかけるデータを収集する予定であった。この理由はアンケートの設問項目について説得力を持ったものが設定できる、アンケート作成者による極めて個人的な主観による設問になることを防ぐという目的があった。このアンケートはのちに述べるとおりに実施を断念しているが、設問の作成まではおこなった。以下にその内容を記す。設問項目は

- 「所属コースはどこか」
- 「学年は」
- 「性別は」
- 「講義以外で精力的に取り組んでいると思うものを選択してください」
- 「貴方がコース振り分け後に身についたと思う技能について選択してください」
- 「自身がこれまで作成した成果物（講義内容は問いません）について印象的なものを記述してください。(複数回答可)」
- 「履修登録の際の動機と考えるものを選択してください。」
- 「あなたがコースの講義の中で特徴的だと思うものについて記述してください。」

このうち「講義以外で精力的に取り組んでいると思うものを選択してください」、「貴方がコース振り分け後に身についたと思う技能について選択してください」、「履修登録の際の動機と考えるものを選択してください。」では選択肢からの複数回答可とした。それぞれの選択肢の内容は以下のとおりである。

「講義以外で精力的に取り組んでいると思うものを選択してください」では、

- バイト
- ボランティア

- 音楽鑑賞
- ゲーム
- サークル
- スポーツ
- 映画鑑賞
- TV 鑑賞
- 読書
- 旅行
- その他

「貴方がコース振り分け後に身についたと思う技能について選択してください」では、

- 物理学
- 情報リテラシー
- 英語
- アルゴリズム的思考
- 電子工学
- プレゼン能力
- レイアウト
- レポート作成能力
- プログラミング
- ソフトウェア知識
- ハードウェア知識
- デッサン
- ネットワーク知識
- データベース知識
- 生物学
- 数学
- 経済
- その他

「履修登録の際の動機と考えるものを選択してください。」では、

- 就職後などの将来を見据えて
- 講義内容に興味が持てたため
- 担当教員に興味が持てたため
- 他の履修している講義と関連があるため
- その他

となっていた。「自身がこれまで作成した成果物（講義内容は問いません）について印象的なものを記述してください。(複数回答可)」では成果物の説明とその理由を記述方式で、「あなたがコースの講義の中で特徴的だと思うものについて記述してください。」では講義名とその理由を記述方式で、それぞれ問う設問であった。

また、これを断念した理由は同じグループから複数回アンケート協力の依頼があった場合に二回目以降の回答に杜撰なものが現れるのではないかという懸念を担当教員から指摘されたためであ

る。この懸念をアンケート担当グループで話し合った結果、限られた期間において有意義なアンケートを収集するためには複数回アンケート実施するべきではないとし、アンケートは機械学習に必要なデータのためのものに限った一回のみにするという方針に変更した。

設問提案

設問の作成にあたってはアンケート担当グループ内で案を出し合った。まず、各コースの特徴を出すことができる設問にするためのキーワードはなにかを検討した。その際中間発表においてデモンストレーションをするにあたって収集したアンケートではどのような設問であればコースごとの特徴が検出できたか、講義のシラバスや当学のウェブホームページやパンフレットに掲示されているコース紹介ではどのようなキーワードが挙げられているか、自身の体験においてどのようなことによりコースごとの特徴が出ているかを振り返ることを参考にした。次に、そのキーワードを元にコースごとの特徴が出るのではないかと思われる設問を作成していった。

設問検討

節で作成した設問のうちから実際に実施するアンケートに含む項目を選定した。その際に、個人の人格でなく資質を問うような設問であること、回答の際に出来る限り選択式で回答できる設問であることの 2 点を基準とした。個人の人格でなく資質を問うような設問であることとはコース選択支援の際に本人の勉強したい内容、目的に関連しない項目を除外する目的がある。例えば、「どのようなサークル活動をしているか」、「どのような余暇の過ごし方を望むか」等は直接学問に関連するものではなくコース選択支援においてこの設問で判別することは適切ではないと判断したかったからである。後者は回答者の負担を減らすことによって出来る限りの収集しやすさと回答者の正直な解答を引き出すことを目的としたものである。ただし、記述式の設問は回答者ごとに大きな特徴が出て判別に役に立つであろうことが予測できたため設問として採用することを可能とした。

設問形式検討

アンケートを実施するにあたって設問形式をどのような形で実施するかを検討した。検討内容は主に回答者の負担を減らすことを目的とした検討となる。具体的には出来る限り記述ではなく選択による回答を求めること、記述の設問を設定する際には記述がしやすいようにすることを考えた。選択式の設問では設問にたいして自身について 5 段階評価で当てはまるものを選んでもらう形にし、記述の設問では回答の参考になるようなキーワードを併記する工夫をして質問項目を設定した。その結果選択式の設問と記述を求める設問、加えて性別、学年、コースを問う設問をあわせて実施するアンケートの設問とした。実施したアンケートの項目は以下のとおりである。

- 「所属コース」
- 「性別」
- 「学年」
- 「システム開発に係る技術（銀行システムや、航空機の予約システム作成など）に興味はありますか？」
- 「web デザインに興味はありますか？」
- 「画像や映像を編集する作業をするのは好きですか？」
- 「ロボットに興味はありますか？」
- 「人工知能に興味はありますか？」
- 「経済に興味はありますか？」

- 「ハードウェア技術に興味はありますか？」
- 「認知科学に興味はありますか？」
- 「電子工学に興味はありますか？」
- 「数学に興味はありますか？」
- 「パソコンやスマホなどを使っているとき、使いやすさやその改善点などを意識していますか？」
- 「人とは違う学問を学びたいと思いますか？」
- 「ものづくりを学びたい（学んだ）と思いますか？」
- 「機械を操作するより中身の方が興味はありますか？（ソフトウェア、ハードウェア、OS)」
- 「モノを進化させるとき、”性能・機能の向上”と”使いやすさの向上”ではどちらを優先しますか？」
- 「コミュニケーションを取ったり、プレゼンテーションを行うのが好きですか？」
- 「できるだけ広く様々なことを学びたいですか？」
- 「プログラム開発環境に興味はありますか？」
- 「取得した資格をご記入ください。その他に記入する場合は複数回答可です。」
- 「所属しているコースで何を学んだかご記入ください。」

このうち「所属コース」「性別」「学年」は機械学習のラベル付のための情報として収集し、「モノを進化させるとき、”性能・機能の向上”と”使いやすさの向上”ではどちらを優先しますか？」では二択での回答を求めた。「取得した資格をご記入ください。その他に記入する場合は複数回答可です。」では『IT パスポート』『基本情報技術者』『応用情報技術者』『取得した資格はない』に加えて複数回答可の記述として『その他』を設けてそれらから複数の回答を選択できる形式にした。これらの設問と「所属しているコースで何を学んだかご記入ください。」以外の項目では 5 段階評価の設問とした。また、「所属しているコースで何を学んだかご記入ください。」については以下のキーワードを併記した。

- 数学（カオス・フラクタル、線形代数、微分積分など）
- 確率・統計
- ロボット (人工知能、筋電義手など)
- 情報通信技術 (携帯の通信とか無線 LAN とか)
- メディア (情報伝達媒体)
- 生物学（生物システム、ブレインサイエンスなど）
- ユーザインターフェイス
- ソフトウェア開発
- 保守
- 画像処理
- 音声処理
- 力学
- 経済
- ハードウェア開発
- web デザイン
- データの管理
- 認知心理学

- 情報
- マネージメント（企業戦略など）
- ネットワーク
- サーバーの管理
- アプリ開発
- 調査データ解析 (質問とか実験の)
- 制御

回答者にはこれらのキーワードを元に記述回答の作成をお願いした。また、このキーワードを使用しない記述回答も有効なものとして取り扱うようにした。

（文責: 山田林太郎）

3.4.3 アンケートページの作成

実施方法

設問形式の議論の後には、実際にアンケートを実施するにあたってどのような媒体でアンケートを作成し、実施するかについての検討を行った。案として浮かんだのは、紙媒体での実施と web ページでの実施である。紙媒体で実施した場合のメリットとしては、仮に協力者が一箇所に集まった空間で協力をお願いする場合に、結果の収集が非常に簡単であることである。一方で、web ページで実施した場合には、パソコン等の電子機器で気軽にアンケートに返答することが可能であると予測できた。しかし、今回アンケートを実施するのは本学の 2 年生から 4 年生の学生であり、基本的に全員がパソコンを所持していることや、実施時の状況等を踏まえて web ページでアンケートに答えてもらうことで、より望ましい結果が得られると推測し、そこで我々はアンケートページでの実施を行うこととした。

ページの作成

アンケートページの作成については、google フォームを用いて作成した。設問形式については 1.4 節で述べたように記述の設問と、選択式の設問である。そこで我々は、どのようなページであれば最もアンケートの集計が、より正しい結果で得られるかを検討した。利用者の立場になって考えると、やはりアンケートに答えるというのは非常に手間のかかることであり、敬遠されがちなのではないかと推測できた。そのような心理状況でアンケートに答えてもらった場合に、最も避けたいのは選択式の設問において、答えてもらう方が特に自分の考えに関係なく、無造作にチェックをされることである。その場合に得られた結果は正しいデータではなく、また、集計の際にこちらでその区別を付けることも不可能に近い。そこで、選択式の設問においては、数字については順当で、5 段階評価のうちの最も高い 5 が右端で、最も低い 1 が左端にくる質問と、その順番が逆で最も高い 5 が左端で、最も低い 1 が右端にくる質問を無造作に分けた形でアンケートを作成し、完成させた。しかし、実際にアンケートの実施を行っていると、質問に答えにくいという声が多々上げられた。そういう意見が上げられるのはある程度推測ができたことであったが、あまりにもその意見が多かったため、選択式の設問においては、最も高い 5 が右端で、最も低い 1 が左端に位置する形式に全て統一し、その後のアンケートを実施した。

（文責: 田中健介）

3.4.4　アンケートの収集

システムを作る際用いた機械学習は大量のデータから特徴を抽出するため、作成時に使用するデータの数と質によって私たちの作成するシステムの精度が大きく左右される。従って、そのデータを収集するアンケート収集は重要度の高い活動であった。

それを行ったデータ収集班ではより質のよく、より多くアンケートを回収できるように創意工夫をした。まずシステム作成用のアンケートを作成するにあたって、コースの特徴を抽出してくれる効果的な質問をするためにコースの特徴とは何かを質問するアンケートを実施しようと試みた。しかし、アンケートに 2 回回答してもらう場合 2 回目のアンケートで数量が減ってしまうという懸念があり、それぞれのコースについてデータ収集班で議論していくこととなった。

アンケートの質に関しては、各コースの特徴がより鮮明に抽出できるように、コースの特徴を表すキーワードをまず考えそれをもとに質問項目を作成した。キーワードの例をあげるとすると、複雑系知能コースの場合「ロボット」、情報システムコースの場合「プログラミング」、複雑系コースの場合「経済・数学」、デザインコースの場合「デザイン」などである。また、アンケートの対象をすでにコース選択し実際に所属している 2〜4 年生とした。

課題として、アンケートを作成する際より良いシステム作成をするには、なるべく多くの質問数が必要となるが、質問数を増やし過ぎてしまった場合、回答者が面倒だと思ってしまい結果的に数が減ってしまう可能性があるということがあった。

そこで、多量の質問項目の候補から議論を行って、コースを選定するにあたって必要そうであると判断した特徴的なものを厳選し、質問数を最小限に抑えた。

結果、232 件（情シス 84 件、デザイン 56 件、知能 50 件、複雑 42 件）（2 年 76 件、3 年 107 件、4 年 49 件）（女性 48 件、男性 184 件）のアンケートを収集することが出来た。

（文責: 岩橋賢吾）

3.4.5　データの収集

作成するコース選択支援システムは、コースの適性度を表す結果の表示のみができるが、それだけではコース選択支援システムとして不十分であると判断した。そこで、コース選択支援システムを補助するページを作成することとし、そのページに掲載するデータを収集した。適性結果と共に４つのコース（情報システムコース、情報デザインコース、複雑系コース、知能システムコース）に関する「各コースの特徴」、「各コースで学べること」、「各コースの就職先について」の３つのデータを収集すること、各コースの卒業研究データを収集する作業を行い、その他に「教員情報」、「講義内容」、「プロジェクト学習情報」、「研究室情報」のデータを収集する作業を行った。

4 つのコースの基本情報

はじめに、４つのコース（情報システムコース、情報デザインコース、複雑系コース、知能システムコース）に関する各コースの特徴についてのデータ収集を行った。まず、データの収集方法、どこからデータを収集するのかを決めるために、データ収集班で議論をした。方法として、公立はこだて未来大学２０１５年度版の大学案内パンフレット、１年次から２年次に進級する際に行うコース選択をするための参考資料として、「情報アーキテクチャ学科　情報システムコース」、「２０１２　コースオリエンテーション資料　情報デザインコース」、「複雑系科学コース紹介」、「知能

システムへの誘い」の４つの PDF 資料、「情報アーキテクチャ学科　情報システムコースのページ」、「情報デザインコースについて」、「複雑系コースのページ」、「知能システムコース」の４つの web ページ、アンケートの自由記述を参考にして、データ収集班がそれぞれ、参考資料を見て、重要と思う単語や、文章を抜き出し、google ドライブでドキュメントを共有し、それぞれ「情報システムコース」、「情報デザインコース」、「複雑系コース」、「知能システムコース」の４つのスペースに分けて、データ収集班がそれぞれ、集めたデータと、参考にした資料を書き込んでいく作業を行った。そこから、被っている単語や、文章を削除していく作業、単語、文書ごとにまとめる作業を行った。

各コースで学べること

次に、４つのコース（情報システムコース、情報デザインコース、複雑系コース、知能システムコース）に関する各コースで学べることについてのデータ収集を行った。データ収集班で議論をした結果、各コースの特徴についてのデータ収集方法と同様に、データを集める場所を、公立はこだて未来大学２０１５年度版の大学案内パンフレット、１年次から２年次に進級する際に行うコース選択をするための参考資料として、「情報アーキテクチャ学科　情報システムコース」、「２０１２　コースオリエンテーション資料　情報デザインコース」、「複雑系科学コース紹介」、「知能システムへの誘い」の４つの PDF 資料、「情報アーキテクチャ学科　情報システムコースのページ」、「情報デザインコースについて」、「複雑系コースのページ」、「知能システムコース」の４つの web ページ、アンケートの自由記述を参考にした。データ収集方法は、データ収集班がそれぞれ、参考資料を見て、重要と思う単語や、文章を抜き出したデータを、google ドライブでドキュメントを共有し、ドキュメントにそれぞれ「情報システムコース」、「情報デザインコース」、「複雑系コース」、「知能システムコース」の４つのスペースを分けて、データ収集班がそれぞれ、集めたデータと、参考にした資料を書き込んでいく作業を行った。そこから、被っている単語や、文章を削除していく作業、単語、文書ごとにまとめる作業を各コースの特徴についてのデータ収集方法と同様に行った。

就職先情報

次に、４つのコース（情報システムコース、情報デザインコース、複雑系コース、知能システムコース）に関する各コースごとの就職先のデータ収集を行うことを決めた。データの収集方法、どこからデータを収集するのかを決めるために、データ収集班で議論をした結果、各コースの特徴についてのデータ収集方法、各コースで学べることについてのデータ収集方法と同様に、データを集める参考資料を、公立はこだて未来大学２０１５年度版の大学案内パンフレット、１年次から２年次に進級する際に行うコース選択をするための参考資料として、「情報アーキテクチャ学科　情報システムコース」、「２０１２　コースオリエンテーション資料　情報デザインコース」、「複雑系科学コース紹介」、「知能システムへの誘い」の４つの PDF 資料、「情報アーキテクチャ学科　情報システムコースのページ」、「情報デザインコースについて」、「複雑系コースのページ」、「知能システムコース」の４つの web ページから集めることにしたが、公立はこだて未来大学２０１５年度版の学校案内パンフレットに載っていた就職先のデータは、公立はこだて未来大学全体での就職先データしか載っていなかったため、各コースごとの就職先のデータを得ることができず、使用することができなかった。そのため、１年次から２年次に進級する際に行うコース選択をするための参考資料として、「情報アーキテクチャ学科　情報システムコース」、「２０１２　コースオリエンテーション資料　情報デザインコース」、「複雑系科学コース紹介」、「知能システムへの誘い」の４

つの PDF 資料を参考にしたのだが、情報デザインコースの就職先についてのデータを集めることができたのだが、情報デザインコース以外の３つのコースの就職先についてのデータは載っていなかったため、データを収集することができなかった。その他に、「情報アーキテクチャ学科　情報システムコースのページ」、「情報デザインコースについて」、「複雑系コースのページ」、「知能システムコース」の４つの web ページを参考にしても、各コースの就職先についてのデータが載っていなかったため、参考にした資料からデータを収集することができなかった。そのため、担当教員に相談したところ事務局にお話をお伺いするのがいいというアドバイスを頂いたので、事務局の方に協力をお願いしたが、各コースごとの就職先についてのデータの受け渡しが約一ヶ月以上かかり、最終発表まで一ヶ月を切っていたため、今回は各コースごとの就職先についてのデータを集めることができず、コース選択支援システムの web ページの方でも使用しないことを決めた。そのため、今後の課題として、各コースごとの就職先についてのデータ収集をすることが決まった。

卒業研究について

次に、卒業研究データを収集するために、データの収集方法、どこからデータを収集するのかを決めるために、データ収集班で議論をした。方法として、まず公立はこだて未来大学情報ライブラリーの web ページ上にある「論文を探す」の「未来大学学位論文」から、平成２０年から平成２５年の６年分を年度別に執筆者名と論文名を Google ドライブで共有したスプレッドシートに書き込み、執筆者名から、学内 web サイトの Harbor View Site の「卒業研究に関する情報」の「研究室配属」から執筆者名を検索にかけ、執筆者のコースを調べ、同様にスプレッドシートに書き込みまとめる作業を行った。

各コースの教員情報

次に、各コースの教員が、どのような研究をしているのかを調査した。教員の顔写真があることで、コース選択を支援するための情報が増えると考え、教員情報に関するデータを収集する作業を行った。実際に各教員の研究室を訪問し、教員に写真の使用の許可を頂くことを考えたが、担当教員のアドバイスで、公式ホームページの公立はこだて未来大学の「教員プロフィール」とコース選択支援システムをリンクさせ、教員情報が見れる作業を行った。　そして、最後にグループ B は議論をした結果、「講義内容」、「プロジェクト学習情報」、「研究室情報ついて」のデータを収集した。方法として、講義内容は、平成２６年度版のシラバスの PDF を参考にし、そこから、講義の名前、概要を抜き出し、データを収集した。プロジェクト学習情報については、２０１４年度プロジェクト学習の web ページから、プロジェクト名と概要を抜き出し、データを収集した。研究室情報については、学内 web サイトの Harbor View Site から、研究テーマと概要を抜き出し、データを収集した。プロジェクト学習とコース選択の関係については、プロジェクトの中に、各コースの人たちが、何人いるのかというデータがないため、使用しないことが決まった。

（文責: 祐川翔斗）

3.4.6　システム補助用のページの作成

システム補助の必要性

コース選択システムを作成する上で必要となる事柄がおおまかに分けて 4 つある。

- 入力

どのような入力を与えるのか

- 判別システム
  どのような手法で構成するのか
- 出力
  どのように結果を表示するのか
- 関連情報
  それぞれのコースの特徴を提示する

この中の「関連情報」について提示する方法として、システムを補助するページを作成することとした。

ページの概要

ページの大まかな情報の区分として、「コース紹介」「講義案内」「教員紹介」「プロジェクト学習」「研究室案内」の 5 項目に分けて情報を掲載することにした。また、学内で公開されているコース案内はもちろんのこと、すでにコースに所属している学生にアンケートをとり、どのようなことが学べるかについて、情報をワードクラウド化して表現することで見やすくまとめた。下記画像 1 枚目では、知能システムコースの例を取り上げている。このような「コース紹介」を行った。また、それぞれのコースごとに受講する講義が異なるため、どのような「講義」があるのか、について 2,3,4 年次と学科別にまとめ、講義案内とした。所属するコースを選択する上で決め手となることとして、どのような教員が在籍しているかということも考えた。それぞれの講義を担当している「担当教員」はどのような研究を専攻しているのか、ということについて、検索しやすいように情報をまとめた。「プロジェクト学習」では、現在活動している 22 個のプロジェクトについてまとめた。1 年次と 2 年次は、それぞれのプロジェクトがどのような活動をしているのかということを、前期末のプロジェクトの中間報告発表会や後期末の成果発表会などで知ることができる。しかし、現在の 3 年次は何を学びたいと思いプロジェクトに携わっているのか、どのような方針で活動をするのかなどを知るため情報は、3 年次のプロジェクト学習が始まる前の配属段階にしか公開がされていない。それらを 1 年次に公開することで、自分が何に興味を持って何を勉強するべきなのか、少しでも自分について考えることができると考えた。3 年次に公開された、各教員の研究室で行われた過去の研究テーマ等の紹介をすることで、より自分の将来像を考えるようになると考えられる。画像の 2 枚目のように、コースごと、教員ごとに、過去に行われた研究テーマの一覧を Web ページにて公開をしている。

図 3.1　HTML ページ

HTML によるページの作成

HTML を使って補助用ページを作っていくにあたって、ファイルの場所を正しく指定するため、まず「相対パス」と「絶対パス」について記述する。作成した Web ページ (HOME.html) を実際に開いてみると URL は、「file:///C:/Users/user/Desktop/HOME/HOME.html」というふうに表記されている。この URL は、PC のデスクトップにあるフォルダ名 HOME のなかに存在する HOME.html であるということを表している。左から特定のフォルダの場所をスラッシュで区切り指定しているのである。このように 1 つの事柄を絶対的に表現しているものを「絶対パス」という。一方の「相対パス」は、この HOME.html を現在地として、Web ページ中に画像や別のフォルダのデータを用いる際に、どこのフォルダに存在するかといった道案内をする際に、特定のフォルダを指定する場合のことをいう。現在、HOME というフォルダに HOME.html は存在し、同時に PICTURE という名前のフォルダにある画像 (gazo.jpg) を使おうとする。その場合、「PICTURE/gazo.jpg」というように相対的に表記がなされるのである。

また、HTML の中では、なにをするにしても「タグ」を宣言し、その事柄を記入しなければならない。HTML の構成として、タイトルや文書の要素といった Web ページの情報そのものを記すタグ`<HEAD></HEAD>`と Web ページ中の文章などの本文を構成するタグ`<BODY></BODY>`が必ずなくてはならない。そしてこれら 2 つのタグの中で、さらにタグを作成し、展開することで Web ページとなっていくのである。今回の Web ページ作成にあたって、なかでも私が特に使用した情報と情報をつなげる方法として、`<a href></a>`というタグを紹介する。これはクリックすることで、特定のページにジャンプさせるためのタグである。`<a href="URL">`ページにジャンプ`</a>`

このように URL を絶対パスまたは相対パスで指定することで、指定したページに飛ぶことができる。自分が保持するテキストファイルだとしても、ブラウザで開くことが可能である。また、`<a href="\#">`ページの上に戻る`</a>` のように、URL と表記されていた部分を”#”とすることで、ページの更新、つまりリダイヤルを行うのだ。さらに、ジャンプしたいページを別ウインドウまたは新しいタブで開かせることも可能である。その場合は、`<a href="javascript:void(0)"`
`"window.open('URL','Web ページのタイトル','width=400,height=400');">`別ウインドウで表示する`</a>`といったようになる。クリックされるページはそのままの状態に保たれるように、ジャンプさせる場所を JavaScript を使い、この場所に留まるという意味である void(0) と指定した。また、window.open() というコマンドを使い、新しく開かせるページの場所と高さと横幅を指定した。

css によるページ作成

こうして HTML だけで Web ページを作成することはできる。しかし、情報を提示する上で気をつけなければならないこととして「見やすさ」があげられる。この問題を解決するためには、CSS を使う必要がある。CSS では、フォントの色やサイズなどの表示スタイルを区別したい段落ごとに分けて修飾してくれるものである。CSS を使用する場合、主に適用させたい段落やピンポイントな部分を独立させるタグ`<div></div>`を用いる。さらに独立させる際には、それぞれに要素名を”id”や”class”として定める。この要素名を”p”とした場合、`<div id="p">`本文`</div>`といったように表記できる。そうして、外部ファイルとして CSS ファイルを作り、そのファイルのなかに”p”をどのように修飾するのか、宣言するのである。例として CSS ファイルには、`p {font-size:large; color:blue; line-height:1.5;}`と記入する。こうした場合、文字サイズは通常より大きくし、色は青く、行の高さを 1.5 倍に設定されるのだ。CSS は Web ページの

作成に必要不可欠な要素であり、見出しや段落分け、トップメニューなどを修飾することで、情報と情報をつなぐ強力な手助けとなった。

出力

次に解決しなければならない問題は、コース判別をした際にどのように結果を表示するべきか、という「出力」の表現方法についてだった。プロジェクト学習の中間報告発表会でこのシステムを発表した際に、「一番重要となるのは結果表示である出力であり、その情報を可視化することによって、意思決定に大きな影響を与えることができるだろう。」という評価をもらった。完成したシステムでは、情報の可視化が行われているかどうか定かでないが、可視化という観点で試した3つのことがある。それは、「ワードクラウド」と「立体的円グラフ」である。

ワードクラウド

当初、「入力」を自由記述式と考えた。そして、得られた文章に対して解析を行い、特徴が高い単語を抽出する。これに機械学習を適用し、コースを判別させる。このことから、使用者に対して「ワードクラウド」を提示し、それらについて複数個のパターンのアドバイスを提供することを考えた。「ワードクラウド」とは、文章中で出現頻度が高い単語を複数選び出し、その頻度に応じた大きさで図示する手法である。また、単語に対して、文字の大きさだけでなく、色、字体、向きに変化をつけ、自動的に並び替えて表現する。これを実現するためには、JavaScript を使い「d3.js」と「d3-clud.js」を適用させ、さらに、「JQUERY」を用いて、入力とコース判別システムと出力すべてとの連携を取る技術が必要であった。「d3.js」を適用することで、グラフィカルな描写が可能となる。主にグラフやデータを視覚的に表現するためのスクリプトであり、うってつけだと考えた。しかし表示された単語と単語が重なって見えなくなる現象が生じた。配置の問題について、JavaScript を用いて様々な関数を適用することで解決可能だった可能性があるが、システムの作成には期日があり、それまでに費やす時間との釣り合いが見合わなかったため断念した。次に、この「d3.js」のデフォルトの設定は英語となっているために起こる書式の問題が生じた。書式を変え日本語を適用させた際に、単語の区切りを判別する機能を失ってしまうのである。例えば「複雑系コースは、工学や数理科学、情報科学などの幅広い領域を学ぶことができる。」このような文章をワードクラウドとしようとした際に、結果として表示されたものは、「複雑系コースは、工学」「や数理科学、情報科学などの幅広い」「領域を学ぶことができる」「。」になる。これがどのように区別されているのかについて把握することができず、改善まで手が及ばなかった。この問題については、システム作成の上で重要だと考えたが、次に検討していた事柄を吟味するために、不採用となった。

グラフの表示

次に試したことは、円グラフなどのグラフを使うことである。グラフを使う上で必要となるスクリプトは多数存在する。先ほどの「d3.js」を使うことで立体的に視覚的にも見応えのあるグラフを表示することが可能だったが、シンプルでそのほかのシステムとの連携を取りやすいスクリプトではないと判断したため採用には至らなかった。その他では「ccchart.js」がある。これは、折れ線、棒、積み上げ、面、円、散布図等いろいろなチャートが手軽に描画できるものである。そしてもう1つ「Chart.js」があった。できることは2つとも変わらないが、なかでも「Chart.js」は、JavaScript の記述方法を学習する時間にあまり多くの時間を割かなくても済むよう、すでにスクリプトの中でコーディング不要で理想的なチャートを出力できるように設定がなされている。初心者

にやさしい簡単で使い勝手がいいものだった。少しだけ JavaScript によって修飾することで、比較的ポップなグラフを表示することもできた。この中でも円グラフを用いて、それぞれのコースがどれほどの適正があるかを表すことで、下記の画像のようなシンプルで見やすい出力結果を表示することができた。

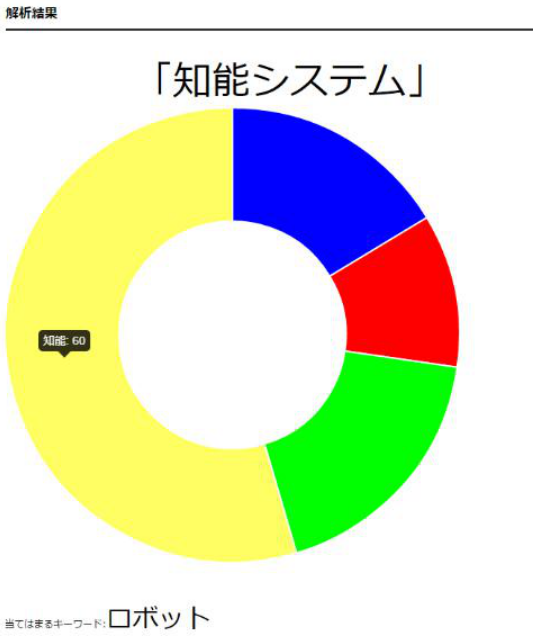

図 3.2　円グラフ

グラフの表示にあたって、2 年次から 4 年次までに「それぞれのコースを表す色を教えて下さい。」と調査をした。その結果では、情報システムコースは「青色」であり、情報デザインコースは「赤色」となり、複雑系コースは「緑色」となり、知能システムコースは「黄色」となり、グラフが栄えあるものになった。完成したシステムを実際に 1 年次に使用してもらった感想では、結果表示が見やすくわかりやすいものという評価であった。　また、これまでのシステムの出力を、どのように入力と連携を取ってグラフに表示をしたのか。このコース選択システムでは、機械学習を Web で行っているわけではない。R 言語のライブラリである Shiny を用いて、下記の画像のような複数の質問に選択式で答える形式を取り、その結果を Web ページとの連携で表示している。

Web ページは、URL という情報を簡略的に記したものを読み取って表示されている。その URL に情報を付加することで、表示されるページにも情報を付加することが可能である。そうするために勉強しなければならなかったのは、「Query の取得」である。我々が作成したシステムにおいて、実際に付加した Query を見てみると、「system.html?course=1&type=1&p1=59&p2=7&p3=21&p4=13」このように表示がされている。URL の末尾に「?」をつけ、それ以降、用意した変数に値を代入し、それを受け渡すのだ。この場合、「course」という変数には 1 という値を入れている。これは、1 ならば情報システムコースを表し、2 ならば情報デザインコースを表す。そして、3 ならば複雑系コースであり、4 ならば知能システムコースである。現在は、「course=1」となっているので、情報システムコースがオススメされているのだ。次に「type」に 1 という値を入れているのがわかる。ここでは、それぞれのコースに適したキーワードをオススメしているのだ。すでにコースに所属している学生に対して調査を行い、何を学びたかったのか、特に特徴のある単語を事前に用意した配列に用意してあるのだ。「type」は、この配列の何番目のキーワードと一致したのかを教えてくれている。実際に用

図 3.3　Shiny ページ

意したキーワードをそれぞれのコースで 5 つあり、それは、情報システムコースならば、”システム”,”管理”,”手法”,”ソフトウェア”,”ネットワーク”となっている。情報デザインコースならば、”デザイン”,”プロセス”,”表現”,”インタフェース”,”設計”である。複雑系コースならば、”複雑”,”力学”,”計算”,”フラクタル”,”科学”とであり、知能システムコースならば、”問題”,”解決”,”認知”,”手法”,”論理”となっている。現在は、「type=1」となっているため、興味関心があるキーワードとして「システム」という単語がグラフとともに表示されている。そして「p1」「p2」「p3」「p4」それぞれに値が与えられている。「p1」は情報システムコースを表す変数名である。これまでと同じように、「p2」は情報デザインコースを表し、「p3」は複雑系コースを表し、「p4」は知能システムコースを表しているのである。そして与えられた値はコースの適性度であるパーセンテージの値であるため、「p1=59&p2=7&p3=21&p4=13」という文字列は、「情報システムコースに 59% 向いており、情報デザインコースは 7%、複雑系コースは 21% であり、知能システムコースは 13% の適正がある」ということになる。また、ここで 6 つの要素をもつ配列「value[5]」を用意した。配列の中には、”適正のあるコースを表す値 (1 4)”,”興味関心のあるキーワードを表す値 (1 4)”,”情報システムコースにおける適正度”,”情報デザインコースにおける適正度”,”複雑系コースにおける適正度”,”知能システムコースにおける適正度”を用意した。Query の取得によって得られた値を保存する配列を用意することによって、グラフやその他の応用にも使用されるのである。しかし JavaScript や HTML の中では、この配列に含まれる数字は文字列であるという認識がなされているのだ。「Chart.js」では、この数字を使うことができずグラフが表示されなかったが、この文字は「数字」である、という変換を行うことで結果の表示が可能となった。今回のようにグラフを扱う場合、結果のグラフが表示されないエラーに苦しんだという事例を耳にするが、そのどれもが下記のようにして、文字列を数値に変換することで解決がされていた。

```
<script>
var p1 = Number(value[2]);
var p2 = Number(value[3]);
var p3 = Number(value[4]);
var p4 = Number(value[5]);
</script>
```

（文責: 中島大貴）

3.4.7　データの分析

目的

本プロジェクトにおけるシステムの概要は、アンケート結果という多変量の変数から機械学習技術を用いて適切なコースを判別するというものである。この判別結果は「結果」しか取り出すことができず、根拠となる部分が弱い。そこで、根拠となるようなデータの法則性を見つけ出すなど、データ分析技術を用いて収集したデータをさらに活用できないか試みることを行った。またこのデータ分析には他にも、システムの構築時に重要度の高いデータを抽出するなども並行して行った。

方法

データ分析には様々な手法があり、データの形や分析対象によって適切な方法を取る必要がある。今回は因子分析という手法を使って分析を進める。因子分析とは、多変量解析と呼ばれる複数の結果変数からなる多変量データを統計的に扱う手法のうちの一種である。測定可能な変数から、その背後にある潜在変数を分析する手法で、心理学や社会学、マーケティングなどに幅広く利用される。本プロジェクトにおいて機械学習技術に R を利用していたが、この因子分析も R を利用して行った。

因子分析の手順

因子分析は、データの背後にある潜在変数を分析することが目的である。多変量のデータから、そのデータの変数を決定付けるような共通の因子を探り、見つけていく。分析の手順は以下の通りである。

1. 因子数の決定
   データを行列として固有値を算出し、固有値の大きい (データの変数に影響の大きい) ものの個数によって因子数におおよその予想を立てて因子数を決定する。ただし、これはおおよその参考程度であり、最終的には試行錯誤を繰り返してちょうど良いものを決定する。
2. 因子分析
   データと因子数を元に因子分析を行う。独立因子や共通因子、寄与率等を参考にして因子数が正しいかどうかを判断する。印指数が正しければ、その因子がなんであるか、どのような共通性が見られるかなどを詳しく分析していく。

サンプルデータを使用した技術習得

因子分析をあまり理解しないままで目的のデータを分析するより、一度わかりやすいデータを使って因子分析の練習をしてからのほうがよいと考えたため、東京図書のサイトにある「因子分析入門」からダウンロードすることができる「6 科目の学力テストデータ」を使って練習を行うこととした。データの一例を以下に示す。

上記の例では上から 5 件ほどの例を表示している。このデータをもとに因子分析を行っていく。

初めに、分析するデータの因子数を決定する。データの固有値を計算すると以下の図の通りであった。この図から、3 つ目の固有値以降は影響が少なそうであると判断できるので、因子数を 2 と定める。

次に因子分析を行う。R では因子分析を行うための関数 factanal が用意されているので、それ

	英語	現代文	古典	数学	物理	地学
1	31	31	33	59	63	52
2	72	77	60	50	51	68
3	34	27	47	34	33	20
4	44	44	54	43	35	48
5	58	57	54	45	61	63

表 3.6　6 科目の学力テストデータ例

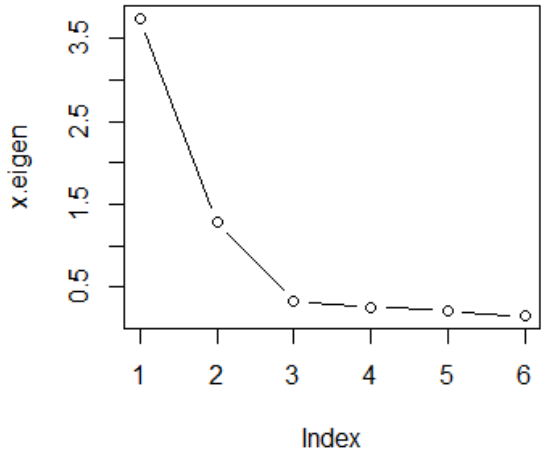

図 3.4　6 科目学力 (量的) 固有値 plot 結果

を使って因子分析を行う。まず、独立因子が以下の通りであった。独立因子とは、共通因子と関係

英語	現代文	古典	数学	物理	地学
0.249	0.129	0.290	0.174	0.255	0.335

表 3.7　6 科目学力テストデータ　独立因子負荷量

のない変数固有の因子のことである。たとえば英語は 0.249、おおよそ 25% 程度は共通因子に含まれない部分であるということを示す。次に共通因子が以下の通りであった。また、これらを棒グラ

	Factor1	Factor2
英語	0.823	
現代文	0.947	
古典	0.869	
数学		0.938
物理		0.865
地学		0.762

表 3.8　6 科目学力テストデータ　共通因子負荷量

フにしたものが以下の図である。

第一因子 (Factor1) は英語と現代文と古典の負荷が高く、第二因子 (Factor2) は数学、地理、地学の負荷が高くなっていることがわかる。ここから、科目の関係性を考えておよそ「理系科目」と「文系科目」という 2 つの因子が浮かび上がる様子がわかる。

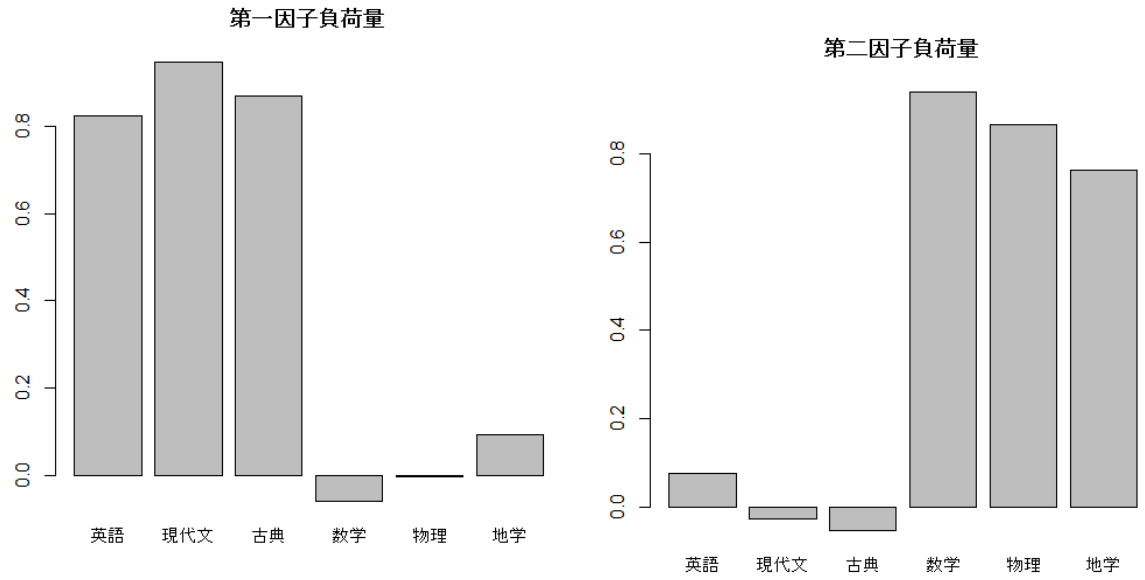

図 3.5　6 科目学力テスト　因子負荷量

また、これらの結果とともに統計的に p 値が算出される。今回の p 値は 1.2^{-13} となるので、この結果は有意であると判断できる。

最後に、データの可視化を目的としたいため、これらの結果をひとつの図にまとめる biplot というプロット方法を使って散布図を描く。以下にその図を示す。この図からも第一因子と第二因子

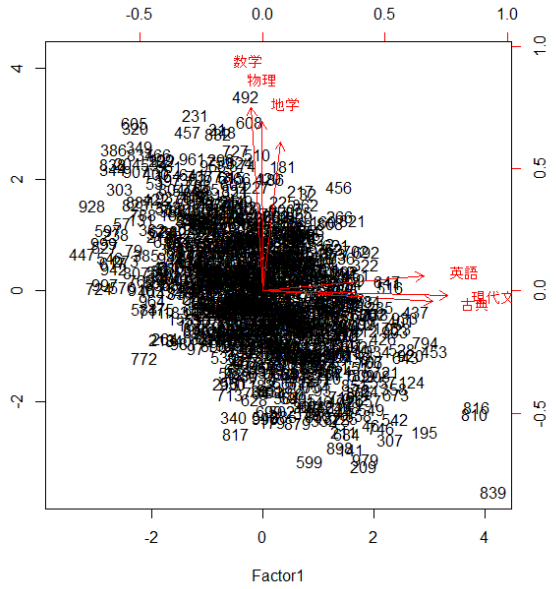

図 3.6　6 科目学力テスト　散布図

の関係性が見られる。黒字で書かれた数値は個人がこの散布図のどこに位置するかを表す。全体で見ると、やや右下がりの楕円形になっている。これは、「理系因子を持つが文系因子を持たない (左上)」「文系因子を持つが理系因子を持たない (右下)」という層ができていることが伺える。同時に、「どちらの因子も持つ (右上)」「どちらの因子も持たない (左下)」という層が薄いこともわかる。

重要項目の抽出

実際に分析する対象のデータはアンケート項目に対する回答であり、19 の項目があった。以下に 19 の項目を示す。

1. システム開発に関わる技術（銀行システムや、航空機の予約システム作成など）に興味はありますか？
2. web デザインに興味はありますか？
3. 画像や映像を編集する作業をするのは好きですか？
4. ロボットに興味はありますか？
5. 人工知能に興味はありますか？
6. 経済に興味はありますか？
7. ハードウェア技術に興味はありますか？
8. 認知科学に興味はありますか？
9. 電子工学に興味はありますか？
10. 数学に興味はありますか？
11. パソコンやスマホなど使っているとき、使いやすさやその改善点などを意識していますか？
12. 人とは違う学問を学びたいと思いますか？
13. ものづくりを学びたい（学んだ）と思いますか？
14. 機械を操作するより中身の方が興味はありますか？（ソフトウェア、ハードウェア、ＯＳ）
15. モノを進化させるとき、”性能・機能の向上”と”使いやすさ”の向上とではどちらを優先しますか？
16. コミュニケーションを取ったり、プレゼンテーションを行うのが好きですか？
17. できるだけ広く様々なことを学びたいですか？
18. プログラム開発環境に興味はありますか？
19. 取得した資格をご記入ください。その他に記入する場合は複数回答可です。

この 19 項目には、機械学習による判別に必要ではない要素が含まれる可能性がある。その必要ではない要素と、必要な要素を区別するために先ほどの因子分析を行った。

これらの項目に対する回答から、先ほどの手順と同様にして因子分析を行う。第四因子までを分析したものが以下の図である。図にあるように、特に重要性の高そうな質問を 10 個ほど抽出してシステムテストを行ったところ、19 の項目での学習よりも 10 の項目での学習のほうがよりよい結果を出すことがわかった。おそらく、前述のような機械学習に必要のない要素を省けたことが精度の上昇につながったと考えられる。そのため、システムに使用する質問項目と、データの分析に使う質問項目は以下の 10 個の項目とした。

- 1. システム開発に関わる技術（銀行システムや、航空機の予約システム作成など）に興味はありますか？
- 2.web デザインに興味はありますか？
- 4. ロボットに興味はありますか？
- 6. 経済に興味はありますか？
- 7. ハードウェア技術に興味はありますか？
- 8. 認知科学に興味はありますか？
- 9. 電子工学に興味はありますか？

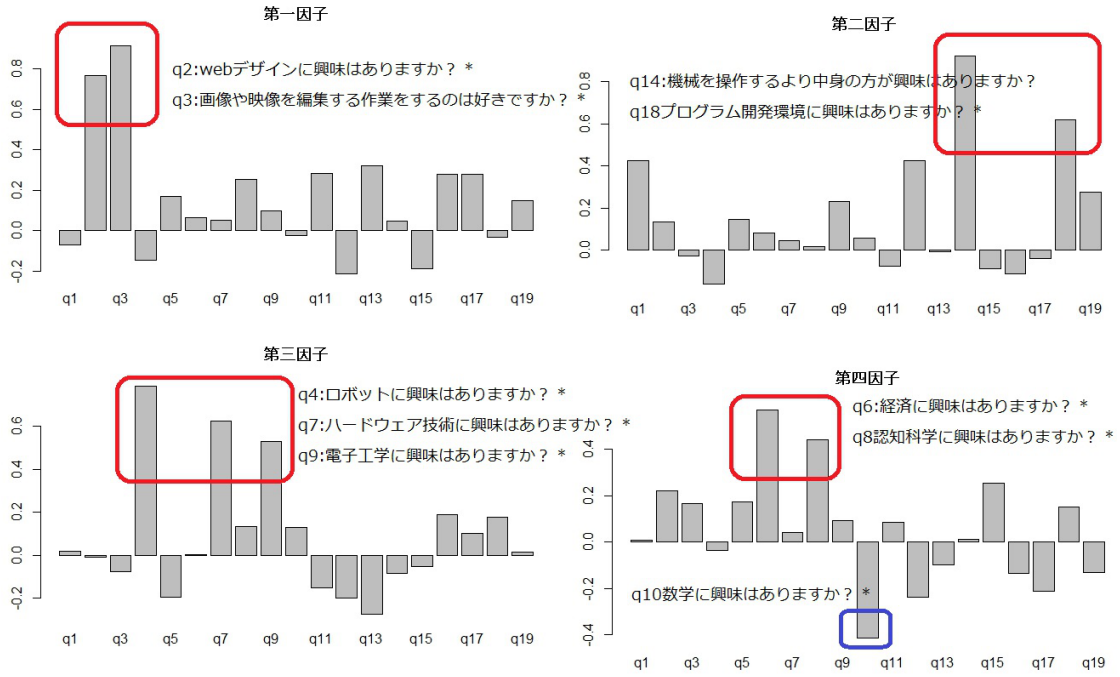

図 3.7　19 項目因子分析結果

- 11. パソコンやスマホなど使っているとき、使いやすさやその改善点などを意識していますか？
- 14. 機械を操作するより中身の方が興味はありますか？（ソフトウェア、ハードウェア、ＯＳ）
- 18. プログラム開発環境に興味はありますか？

因子負荷量に対するデータの分析

次に、上記の 10 の質問項目のみの回答を抽出し、それらを因子分析によってさらに詳しく分析を進めた。まず、因子数を決定するにあたって、3 つ前後というおおよそのあたりをつけて、p 値より妥当性から適切なものを検討した。因子数が 2 のときは p 値が 0.034、因子数が 3 のときは p 値が 0.479、因子数が 4 のときは p 値が 0.972 であった。これらの試行錯誤の結果、因子数は 2 が妥当であると判断した。因子分析の結果が以下の通りである。

この結果を棒グラフに表すと次のようになる。図の第一因子を見ると、「ロボットに興味はありますか？」「ハードウェア技術に興味はありますか？」「電子工学に興味はありますか？」という項目に強い共通性が見られる。これは内容から推測すると、「機械技術への関心の因子」と思われる。次に、第二因子を見ると、「システム開発に関わる技術（銀行システムや、航空機の予約システム作成など）に興味はありますか？」「機械を操作するより中身の方が興味はありますか？（ソフトウェア、ハードウェア、ＯＳ）」「プログラム開発環境に興味はありますか？」という項目に強い共通性が見られる。これも同様に推測すると、「ソフトウェア技術への関心の因子」と思われる。逆にこの第二因子には負の負荷量を示している項目が「経済に興味はありますか？」「認知科学に興味はありますか？」というもの。これより、ソフトウェア技術への関心因子を持つ人は、経済や認知科学等の利用法への関心が薄い可能性が見られる。

	Factor1	Factor2
q1		0.473
q2		
q4	0.547	
q6	0.223	-0.347
q7	0.777	0.113
q8	0.151	-0.252
q9	0.654	
q11		
q14	0.154	0.466
q18	0.233	0.644

表 3.9　アンケートデータ　共通因子負荷量

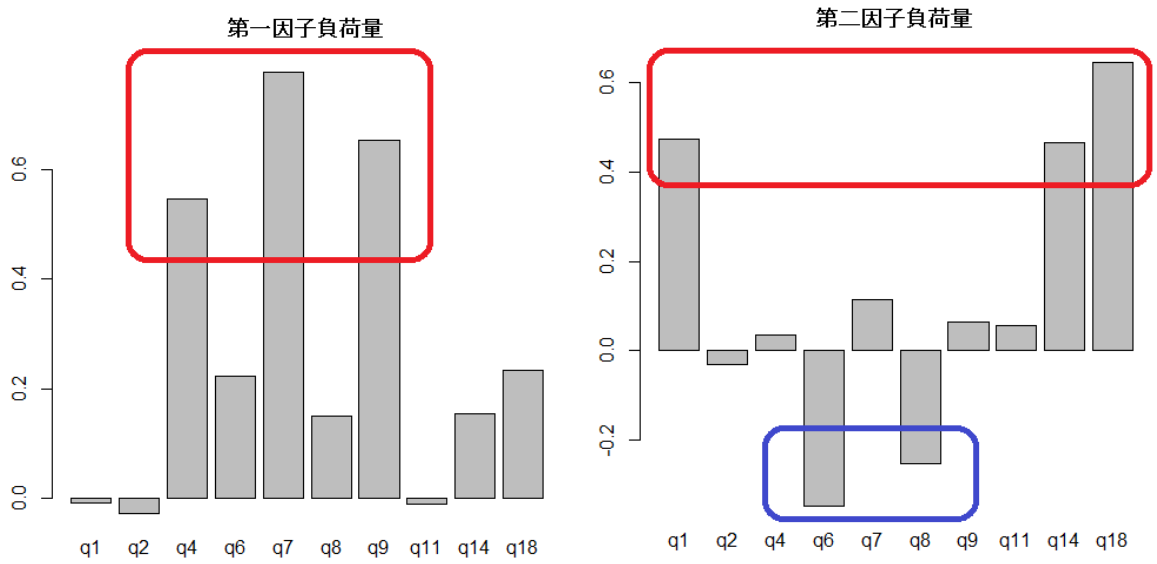

図 3.8　アンケート回答結果に対する因子負荷量

biplot を用いたデータ分析

最後に、biplot を用いてデータを分析した。biplot の図への表示は回答者のコースごとに色分けを行った。また、全ての回答を一度にプロットすると非常に見にくいため、4 つにわけて表示した。以下に図を示す。矢印はそれぞれ分析した質問項目の因子付加量を示している。先ほどの第二因子で正反対の負荷量を示していた項目もそれぞれの逆側に位置していることがわかる。このプロット図から、4 つのコースに何かの法則性が見えるのではないかと考えていたが、あまり重要な手がかりは得られなかった。デザインコースのプロット図はかろうじて「第二因子の絶対値が低い」ため、全体的にやや水平に分布していることがわかる。また、情報システムコースのプロット図は、第二因子の正の方向にやや偏って分布していて、おそらくは「機械技術への関心の因子」が少し強く出ているものと思われる。

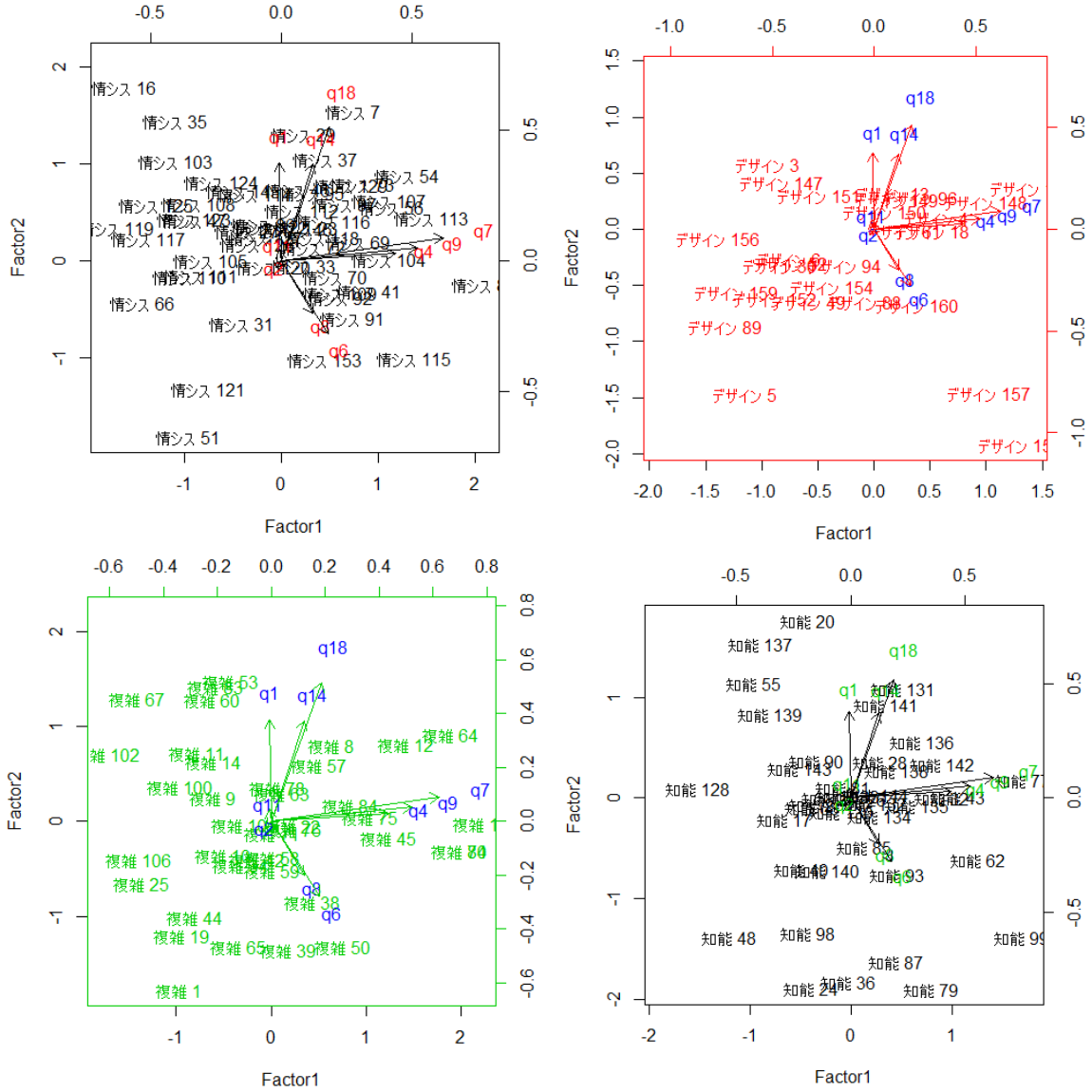

図 3.9　因子分析したデータの biplot 図

結果

アンケートの項目数の絞込みはほぼ成功したといってよい。しかし、その後に行ったデータの分析はうまくいかずに、成果を出すことができなかった。予想として、アンケートの回答データから 4 つのコースの特徴や因子のようなものが抽出することができるのではないかと考えてのデータ分析だったが、結果はうまくいかなかった。そのため、システムにこの分析内容を活用することはなかった。

（文責: 井川翼）

3.4.8 システムの作成

データの形式

アンケートへの回答によって得られたデータが蓄積されている csv ファイルの内容は、日本語の文字列が含まれている部分があり、そのままでは R 言語で数値計算を通して行われる解析手法を適用させていくのには適さない形式となっているので、データの解析処理を行う前にデータ内容を数値データとして置き換える必要がある。この、アンケート回答結果データを R 言語での処理に適する形式に加工して体裁を整える作業を以下の手順で行った。

1. 解析処理に用いない質問項目に関する部分や、データが欠損している部分を取り除く
2. 所属コースを答える質問に関する部分をラベル情報として切り出す。
3. その他の質問で、実際のアンケートでの質問順と異なる順番で記録されている質問項目に関する部分を入れ替えて、アンケートの体裁に合わせる
4. 質問項目名が記入されているデータの列名を、所属コースを答える質問の列名を”class” と、その他の質問の列名を”q1”, ”q2”, …, ”q10”とする
5. 所属コースを答える質問への回答データの文字列を、”情報システム”を”1”に、”情報デザイン”を”2”に、”複雑系”を”3”に、”知能システム”を”4”に置換する

以下に、加工したデータの一例を示す。

q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	class
4	2	4	5	4	2	2	3	4	4	1
5	3	4	2	5	3	2	5	5	5	1
5	1	1	3	1	3	1	4	5	5	1
3	2	4	3	1	4	1	1	2	5	2
3	4	4	2	4	4	2	1	2	4	2
1	3	1	5	1	2	1	4	1	1	2
1	5	1	5	1	5	1	3	1	1	3
4	1	4	1	4	4	3	2	4	4	3
4	4	3	3	2	1	2	3	2	3	3
3	4	2	5	2	4	2	4	2	4	4
2	4	4	3	4	5	4	3	2	4	4
2	2	3	2	2	4	1	4	2	3	4

表 3.10　加工データの一例

これを訓練用データとして、データ解析していく。

機械学習手法の検討

今回のシステムを実現させるために必要となる、データ群から特徴のある部分を抜き出す処理や、目標となるラベルの値を推測する処理に関しては、既存で様々な手法が公表されていて、R 言語で実装ができるようにパッケージとして公開されている。システムの実装を行う前に、まずはシステムで必要となる各処理に対して、それぞれどのような手法を利用していくのが適切であるかを、解析手法に関する調査と検討を重ねながら判断していった。

推薦コースの選定に用いる手法

システムの中心的な出力となる、推薦するコースを決めるための判別システムに使用する機械学習の手法について検討を行い、その結果以下の 3 つの手法が候補として挙げられた。

- LDA(線形判別分析)
- SVM(サポートベクターマシン)
- ランダムフォレスト

そして、これらの手法の中からひとつに絞るため、以下の手順で判別システムを作成した時の分類精度を予測した。

1. アンケートから得られた結果から、8 割を判別システムを作成する時の訓練データ、2 割をシステムの分類の実験に用いるテストデータとして、ランダムにサンプルする。
2. 訓練データからそれぞれの手法で判別システムを作成し、テストデータに対してシステムが予測した所属コースを示すデータのラベルが、実際のデータのラベルと一致しているかを調べていく。
3. 全てのテストデータで調べ、正しくラベルの予測が出来ていた割合を算出する。
4. 1 3 を 5 回繰り返して、5 回分の算出された割合の平均値をシステムの分類精度とする。

分類精度の予測の結果、LDA は約 42%、SVM は約 84%、ランダムフォレストは約 40% の精度で分類に成功していたので、今回のシステムでの推薦コースの選定には SVM を用いることに決めた。

キーワードの選定に用いる手法

システムの補助的な出力となる、推薦されたコースに関わるキーワードの中で、特にどのキーワードに関心が強そうかを予測するための手法としては、調査の結果、因子分析と呼ばれる解析手法を用いることに決定した。

システムの実装

アンケートへの回答によって得られたデータと、調査した機械学習の手法、データの解析手法を用いて、R 言語を利用して実際にコース選択支援システムの実装に取り組んでいった。

判別システムの実装

システムの中心的な出力となる、推薦するコースを決めるための判別システムを、アンケートから得られた結果を元に、各質問への回答内容から所属コースを表すラベルの値を予測するような SVM の予測モデルを作成して、オブジェクトをファイルとして出力した。また、目安として各コースへの適正度の数値も出力するため、目的となるコースに所属してる学生のデータのラベルを 1、それ以外のデータのラベルを 0 として、ラベルの値を量的に予測させる予測モデルも作成して、オブジェクトをファイルとして出力した。

キーワードの決定

システムの補助的な出力となる、推薦されたコースに関わるキーワードの中で、特にどのキーワードに関心が強そうかを予測する機能を作成するため、まずは各コースに対するキーワードを決

定するための技術的な根拠のある候補を出していくために、本学で行われる 2〜4 年生を対象とした講義の講義内容と講義を担当する教員名、講義の対象コースの情報を収集して、以下の手順で解析を行っていった。

1. 講義が対象としているコース情報を抜き出し、データを 4 コース分に分ける。
2. 各コースの講義内容の文章に対して形態素解析と単語への重み付けを行い、その結果から名詞のみを抜き出す。
3. 各コースで、各単語に対してコースを特徴的に表しているかを示す、重みの数値を算出し、数値が大きい 30 番目までの単語をキーワードの候補として抽出する。

以下に、解析結果の一例として、各コースでキーワードの候補となった上位 5 単語を示す。

	1	2	3	4	5
情報システム	技術	的	開発	システム	手法
情報デザイン	デザイン	人	情報	プロセス	表現
複雑系	的	力学	複雑	基礎	学
知能システム	問題	協調	解決	方程式	制御

表 3.11　キーワード上位 5 つ

また、過年度の本学学生の卒業論文タイトルの情報を収集して、上記と同じ手順で解析を行い、キーワードの候補を抽出した後、2 つの抽出結果を参考に、話し合いでシステムに利用するキーワードを決定した。　以下に決定したキーワードを示す。

	1	2	3	4	5
情報システム	システム	管理	手法	ソフトウェア	ネットワーク
情報デザイン	デザイン	プロセス	表現	インタフェース	設計
複雑系	複雑力	学	計算	フラクタル	科学
知能システム	問題	解決	認知	手法	論理

表 3.12　決定したキーワード

教員の担当講義内容に対するクラスタリング

キーワードの候補を抽出する際に利用した講義に関するデータを利用して、教員の担当講義内容を 4 つのコースに分類することができないかを調べるため、以下の手順で非階層クラスター分析を行った。

1. 講義の担当教員情報を抜き出し、データを教員ごとに分ける。
2. 各教員の担当講義内容に対して、クラスターの数を 4 とした非階層的クラスタリングを行う。
3. 各教員に与えられたクラスターの分類結果となる数値をクラスターの情報として抽出する。

以下に、抽出された結果の一部を示す。

しかし、抽出された情報はほぼ全ての教員が 1 つのクラスターに偏った結果となっていて、この解析結果から参考となる情報は得られないと判断した。他に、講義情報と教員情報だけでなく、卒業論文との情報とあわせてクラスタリングするという案もあったが、情報の収集に手間と時間が

教員	クラスター
Ian Frank	4
片桐恭弘	4
神谷年洋	3
加藤浩仁	4
岡本誠	4
上野嘉夫	4
永野清仁	4
佐藤仁樹	1
竹之内高志	4
V. Riabov	2

表 3.13　各教員と分類クラスター

ある程度かかり、この時点で最終発表までの時間もなかったため、この案は断念した。

因子負荷量の算出

推薦されたコースに関わるキーワードの中で、特にどのキーワードに関心が強そうかを予測する機能を、アンケートへの回答によって得られたデータと決定したキーワードを元に、以下の手順で実装していった。

1. アンケートから得られた結果を学生の所属コース別に分類する。
2. 分類したデータに対して、因子数をキーワード数である 5 個として、因子分析を行う。
3. 因子分析によって得られた結果の中から、データの各項目に対する因子負荷量を示す部分を抜き出す。
4. 因子負荷量の内容から検討して、各キーワードを各因子に対応付けて仮定させる。

以下に、因子負荷量の情報の一例として、情報システムコースの情報に対する因子負荷量を示す。

	Factor1	Factor2	Factor3	Factor4	Factor5
q1	0.41	0.05	0.00	0.01	0.02
q2	-0.11	0.99	0.03	0.05	0.00
q3	0.17	0.04	0.20	0.96	0.06
q4	-0.07	-0.13	0.03	0.09	0.50
q5	0.44	-0.01	0.33	0.24	0.48
q6	0.09	0.36	0.03	-0.13	0.67
q7	0.17	0.02	0.95	0.23	0.07
q8	0.16	-0.10	-0.13	0.09	-0.13
q9	0.58	-0.07	0.03	0.05	0.02
q10	0.65	-0.16	0.26	0.21	-0.10

表 3.14　情報システムコースの因子負荷量

こうして得られた、4 コース分の各質問項目に対する因子負荷量の情報を持つオブジェクトをファイルとして出力した。

web アプリケーションの実装方法

作成したシステムのオブジェクトデータを元に、実際にブラウザ上から質問に回答していくことによって推薦コースやキーワードが表示される web アプリケーションの形式としてシステムを実現させていくため、実現が可能そうな方法について、調査を行った結果、R 言語内の shiny パッケージを使用することに決定した。

UI 部分の実装

システムの UI としては、shiny の特徴を調べて話し合った結果、質問への回答を入力するページと診断結果を表示するページを分けて扱うことに決定した。そのため shiny 上でのページの UI は、アンケートと同内容、同形式の質問を入力するように入力フォームを用意して、サーバ側で算出されたパラメータ値をクエリとして付加させた URL へのリンクを出力することで、web アプリケーション内での動作と結果を表示する HTML ページの動作を切り離すようにした。

サーバ部分の実装

システムの実装で作成されたオブジェクトのファイルを用いながら、web アプリケーションが以下の手順で動作をするように実装を行った。

1. UI 側で入力されたデータを、SVM による予測が適用できるように、アンケートから得られたデータの形式と一致するように加工する。
2. ファイルから読み込んだ SVM オブジェクトデータを用いて、入力されたデータに対する、所属コースを示すラベルの値を予測する。
3. ファイルから読み込んだ SVM オブジェクトデータを用いて、入力されたデータに対する、各コースへの適正度の数値を算出する。
4. 入力されたデータに対するラベルの予測値を推薦するコースと判断して、そのコースに関する関心が強そうなキーワードを予測するため、推薦するコースに対応するキーワードを以下の手順で選出する。
   （a）自由記述式の質問内で、ひとつだけ一致して入力されているキーワードがあった場合、そのキーワードを選出する。
   （b）自由記述式の質問内で、複数の一致して入力されているキーワードがあった場合、ファイルから読み込んだ因子負荷量のデータを用いて、入力データを構成している因子の量を計算し、一致して入力されていたキーワードに対応する因子の中で、構成している因子の量が一番大きい因子に対応しているキーワードを選出する。
   （c）自由記述式の質問内で、一致して入力されているキーワードがなかった場合、ファイルから読み込んだ因子負荷量のデータを用いて、入力データを構成している因子の量を計算し、因子の量が一番大きい因子に対応しているキーワードを選出する。
5. これまでの処理から得られた、推薦するコース、選出するキーワード、各コースに対する適正度を表す数値を、出力ページの URL の文字列にクエリの形式で結合し、その情報を UI 側に送る。

web アプリケーションの稼動

web アプリケーションの UI について記述したファイル、サーバの動作について記述したファイル、サーバ側のファイルで使用する各オブジェクトデータのファイルを同梱して、ローカルまたは web サーバ上で R 言語を起動し、shiny パッケージを読み込んで runApp 関数で実行することにより、システムが稼動する。

（文責: 田中桂介）

第 4 章　成果

4.1　完成したシステム

完成したシステムは次のようなものとなった。R の Shiny というパッケージで、Web ブラウザ

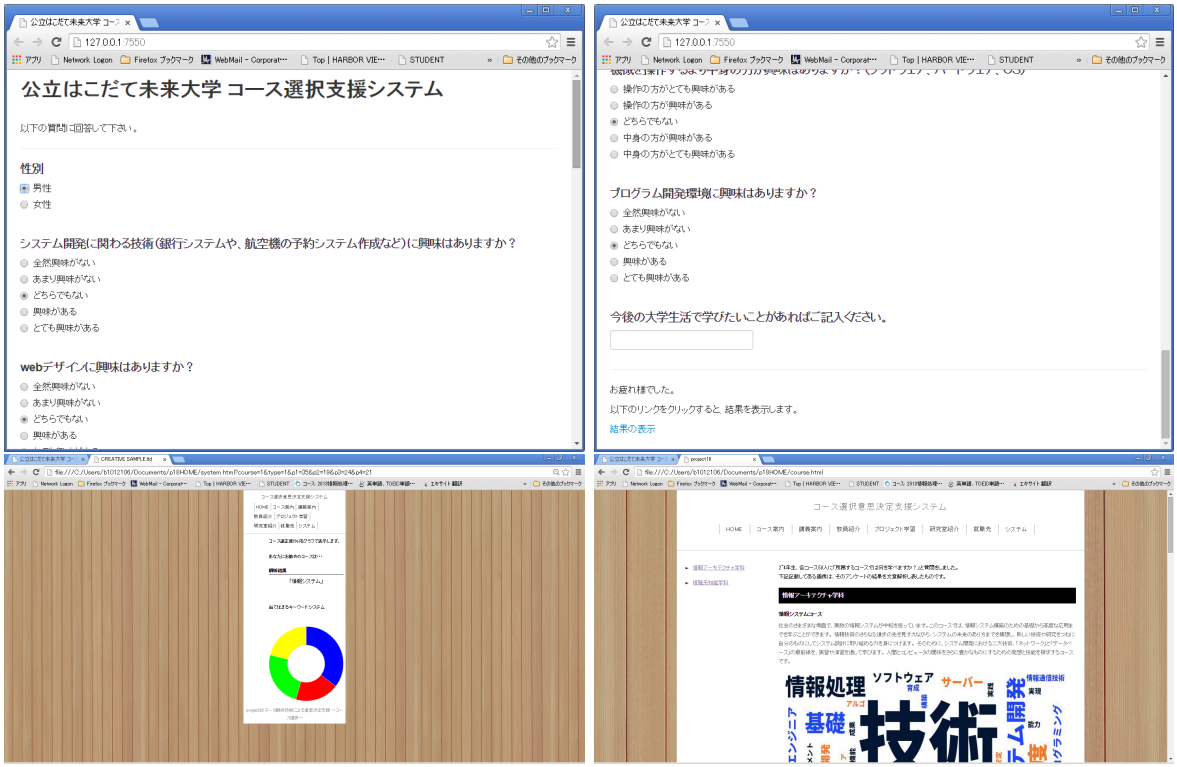

図 4.1　完成したシステム

上でシステムを動作させている。そして質問事項に回答したのちに、HTML ページに飛ぶとシステムの判別した情報がクエリとして引き渡されて、HTML ページにて結果が表示される仕組みである。

（文責: 井川翼）

4.2　システムの評価

1 年生 20 人にシステムを実際に使用してもらい、評価をもらったところ、以下のような結果となった。

1. 参考になった：50%
2. 少し参考になった：20%
3. どちらでもない：20%
4. あまり参考にならなかった：10%

また、最終発表時に多数の意見をもらった。全体を通してもっとも多かったのは、「コースに対する満足度を評価対象に加えてはどうか」というものであった。今回のアンケートでは、コースに所属している人の回答は無条件で「そのコースに適している」という前提で機械学習を行った。しかし、当然だが全員が所属したいコースに所属していたわけではない。また、望んでコースに所属していたとして、そのコースが果たして適切であったのかという問題が生じる。このように、「いかにして所属コースが適切であるか」という問題を解決すべきであったという意見だ。

（文責: 井川翼）

第 5 章　今後の課題と展望

今回はテーマの決定から機械学習についての技術習得、システムの作成とあまり手際よく行かなかったので、今後はもっと計画性をもって事を進めるべきである。また、最終発表時にもらった意見を元に、適切なコース選択ということについてもっと深く研究および議論を深めるべきである。今後の展望について。今回のシステムはすべてローカル上で動作していた。本格的に本学の 1 年生に使用してもらうことを考えると、学内ネットワークにサーバを立ててシステムを構築すべきである。そのため、サーバの構築から行って最終的に本学の学内ネットワークにシステムを実装することが目標となる。

（文責: 井川翼）

参考文献

[1] 石田基広. R によるテキストマイニング入門. 森北出版, 2008.

[2] 読売新聞社, YOMIURI ONLINE, http://www.yomiuri.co.jp/.

[3] 舟尾暢男, R-Tips, http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html.

[4] TAKESHI ARABIKI, R による文書分類入門, http://www.slideshare.net/abicky/r-22325351.

[5] 濱田晃一, R 言語による Random Forest 徹底入門, http://www.slideshare.net/hamadakoichi/introduction-torandomforest-tokyor.

[6] graySpace, 【R によるデータサイエンス】線形判別分析, http://d.hatena.ne.jp/graySpace/20140503/1399106654.

[7] kj-ki, R で nnet を試してみる, http://d.hatena.ne.jp/kj-ki/20120124/p1.