Journal of Culture and Information Science, 11(2), 44-49,（March 2016）

10 周年記念企画

授業実践報告

ビッグデータからの知識発見

―経済・社会データの分析を通じて―

波多野 賢治・宿久 洋・深川 大路

1. はじめに

同志社大学文化情報学部は、人間の営みを文化と捉え、その文化をデータの側面から解析できる人材の育成を目指すために2005 年に設置された。文化情報学部のカリキュラムは、これまで他大学では行われていなかった探究力と思考力を養うために導入され、具体的にはプロジェクト型学習（Project Based Learning: PBL）を実践するために、文化情報学部の講義、演習科目をその準備として位置づけている。このカリキュラムを通じて文化現象を統計解析手法を用いて正確に把握することができる人材、つまり文理の枠を越えたデータサイエンティストの人材育成を図ろうとしている。しかしながら設置から 4 年が経過した時点で、それまでに表面化した問題を修正するために開学部時のカリキュラムを改正し、2009 年度から「ジョイント・リサーチ」と冠する科目（以下、ジョイント・リサーチ系科目）の開講がされた。

ジョイント・リサーチ系科目は、学問探究の基礎を学ぶために参考文献の探索や文献の読み方、データの収集・分析方法、レポート作成方法などをグループ活動を通じて学ぶ。本稿で扱う「ジョイント･リサーチ I」と「ジョイント･リサーチ II」は、その集大成として行われており、あらかじめ答えがあるという保証のない問題に対し、グループ内でさまざまな検討を重ねた結果産み出される創意工夫によって取り組み、問題解決を図っていくという流れで行われている。従来のカリキュラム内でもこのような方針で行われていた科目は存在していたが、扱われてきた文化現象を表すデータ自体のサイズがそれほど大きくはなかった。そのため、いざそれらをデータ分析しようとしても決まり切ったことしか行えず、その結果多角的な分析ができなくなり、各グループでの創意工夫が画一化してしまうという問題が生じていた。このような問題を解消するために設定したテーマが、本稿で取り上げる「ビッグデータからの知識発見」である。

本稿では、この研究テーマで「ジョイント・リサーチ I」「ジョイント・リサーチ II」を運用している著者らのクラスに焦点を当て、世の中の現象を正確に把握し標準化するために必要な能力を涵養するための取組み、つまり、ビッグデータからの知識発見のために必要な知識習得と事例による実習、および経営科学系研究部会連合協議会主催のデータ解析コンペティションへの参加、の効果について報告する。現象を正確に把握して標準化する能力はデータサイエンティストには必須であり、一般的には専門家と長い時間を共に過ごし実践を繰り返すことによって育まれるものであると言われているが、当該クラスでどのような授業展開を行いこうした能力の育成を図っているかについて詳述する。

2. 文化情報学部とビッグデータ

文化情報学部でビッグデータを扱った科目を設置するに至った理由は 1. で述べたとおりであるが、当初はどちらかと言えばデータの量的側面に着目していた点は否めない。しかし、ビッグデータを単にデータの量的側面に着目しただけではその管理にのみ焦点が当てられることになり、文化情報学部のカリキュラムに組み込むべき内容とは言えない。ビッグデータの正確な定義が公式にな

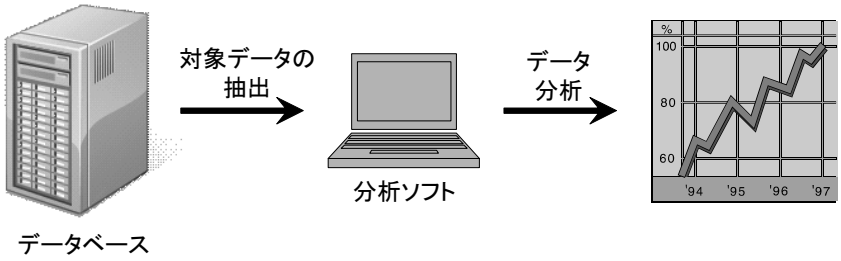

図 1 データ分析の手順

されていないことは周知の事実ではあるが、どの程度のデータ規模かという量的側面の特質だけではなく、どのようなデータから構成されるか、あるいはそのデータがどのように利用されるかという質的側面の特質も併せ持つものであるとされている（総務省、2012）。つまり、対象データ数が多いばかりではなくその変量数も多いため、対象データ間・変量間にさまざまな構造が存在し、一般的な統計解析法の適用は困難なデータである。このようなデータはデータ管理を扱う情報科学だけやデータ分析を扱う統計科学だけではさまざまな問題に対処できないため、学術分野の融合による取組みの中で新たな手法の提案がなされるべきであることを考えると、情報科学と統計科学の専門家がこの問題に取り組む環境が整っている文化情報学部に設置すべきであるといえる。

また文化情報学部ではデータからデータビジネスに活用する知見を引き出す中核人材であるデータサイエンティストの育成がディプロマポリシーの中で挙げられている。これは、これからの企業に求められているものは市場創出力のある新しい商品・サービス・事業・産業の開発であり（古賀、1999）、それを実現するためにはデータサイエンティストの育成が急務であるとされているからである。Reich（1992）は、市場創出力のある新しい商品・サービス・事業・産業の開発は、

- 個々の人間を活用、さらに成長させることができ、それを省資源的に実現できる環境さえあれば、経済は持続成長することができる。
- 経済が持続成長すれば、雇用統計が好転する形で個々の人間がさらに商品やサービス、事業、産業を支えることができる。

という相互補完的な循環を生じさせることができ、このような循環を維持することこそが 21 世紀型資本主義社会に求められていると主張している。そこに多大な貢献をするのが、社会の流行を司る象徴であるデータを分析、操作し、そこから大きな利潤を産み出すシンボリックアナリスト、つまり今で言うデータサイエンティストなのである。

しかし、データサイエンティストを育成すればそれでよいかと言われれば、実はそうではない。上で述べた相互補完的な循環を維持するために、データサイエンティストは自身の得意分野を磨き、また異分野との融合を積極的に行い、問題解決の方法を新たに創発していく必要があるという点は、実は文化情報学部のアドミッションポリシーにも通じるところがある。

以上が 2011 年度から始まった「ジョイント・リサーチ I」、「ジョイント・リサーチ II」でビッグデータを研究テーマに設定するに至った理由である。

3. ジョイント・リサーチ I

「ジョイント・リサーチ I」では、例年、比較的小規模なデータである日経マーケティングリサーチ社の NEED-SCAN/CVS レシートデータ[1]を扱ったデータ分析実習を行っている。

NEED-SCAN/CVS レシートデータは、コンビニエンスストアのレシート 1 枚単位に得られる購入商品の金額や個数に加えて、購入者の年齢・性別・購入日時・場所がデータ分析の対象となるため、同時購買分析や店舗属性・購買者属性・購買日時などによるクロス集計もできる 1 年分のデータとなっているが、受講生の科目履修履歴を考慮[2]し、レシートデータの商品を清涼飲料水とスナック菓子のみに限定している。これは、あまりに大きなデータを情報科学の知識なしに扱おうとすると、データ分析どころではなくなることを危惧したた

[1] http://www.nikkeimm.co.jp/pos/needs-scan-cvs/

[2] 残念なことに文化情報学部は統計学を学ぶ学部だという考えが浸透したためか、情報科学の基礎科目である「情報処理演習」や「アルゴリズムとデータ構造」、「データベースシステム」を履修する学生が少なくなってしまっている。

表 1 「ジョイント・リサーチ I」の授業計画

実施回	内容	授業時間外の学習
1	ガイダンス（講義の目的、実施方法説明）	
2 ～ 4	POS データの分析例と RDB を用いた分析法	講義の復習
5	データの分析手順と分析目標の設定	グループ作業
6 ～ 8	基礎分析（データの傾向把握）	グループ作業
9	中間発表	レポート作成
10 ～ 13	応用分析（基礎分析の結果を活用した取組み）	グループ作業
14、15	最終発表	レポート作成

めである。その結果、扱う必要のあるデータサイズは CSV 形式で 200MB 程度とすることができ、Microsoft Excel でなんとか扱うことができるデータ、つまり情報科学の基礎科目を履修していなくても、また「ジョイント・リサーチ I」の講義内容を理解しづらい場合でも、Microsoft Excel を使ったデータ分析はできるようになっている。

受講生はこのデータを用いて、例えば 2015 年度春学期には、

- 清涼飲料水における商品開発への提唱
- 期間限定商品を売るために
- コンビニエンスストアではどんな商品を何と売ればいいか
- POS データを用いた売上増加研究
- 健康飲料とスナック菓子の同時購買から考えるコンビニの売り上げアップのポイント

をテーマに、各班がデータ分析の結果とその結果を用いた次につながる提案を行っている。このような提案は、一般的に図 1 のような手順を踏んで行われるものだが、2011 年度以前の科目ではデータベースから分析対象データの抽出を行わずに、単に R や IBM SPSS といった分析ソフトを用いて単にデータ分析を行っていたに過ぎない。そのため、現場に即した形式でデータ分析を行うことを目的として、分析前にデータベースを利用して必要なデータ抽出を行うことにしている。

データベースから分析対象データの抽出が行われないことに起因する問題は、分析対象のデータを教員側から与えていた事による分析内容、および方法の画一化であることは既に 1. で述べたとおりである。また 2. で述べたように、社会の流行を司るデータは近年ではビッグデータと呼ばれているため、そのビッグデータを扱えない限りはデータサイエンティストの育成はできないと考えたのである。

「ジョイント・リサーチ I」の具体的な授業計画はおおよそ表 1 に示すとおりである。

- 第 1 回目の講義では、1.、2. で述べた講義の目的を述べた後、この講義の内容が社会のどのような部分で活用されているのかを講述する。
- 第 2 回目の講義では、受講者にとって比較的身近な POS データを用いたデータ分析とその結果を用いた応用事例を紹介し、そうした事例がどのようなプロセスを経て行われるべきなのかを解説する。近年のビッグデータ分析の活用事例は、さまざまな分野が融合された形で行われていることが多いため、身近な POS データを用いた事例を用いて解説している。
- 第 3、4 回目の講義では、分析対象は通常、何も手を加えられていない生データであることを考慮し、生データから分析を始められる状態となるデータベースの構築手順について演習を通じて説明する。NEED-SCAN/CVS レシートデータは、分析対象商品ごとに CSV データとして提供されるため、一枚のレシートデータにするためにはそれぞれの商品のデータを統合しなくてはならない。Microsoft Excel ではそのような処理をすることは非常に難しいが、Microsoft Office Access ではデータ操作言語 SQL を用いれば容易にそうした処理を実現できるので、この 2 回の講義で SQL の便利さを体験してもらうことを特に重要視している。
- 第 5 回目以降の講義では、分析対象のデータがもつ傾向把握のための基礎分析作業、および基礎分析で明らかになったデータの傾向を活用する手法の提案、そしてその提案の実

現に向けた応用分析作業にグループで取り組み、その内容を中間発表と最終発表で報告する。基礎分析で重要な点は、グループメンバ全員でありとあらゆる可能性を考えさまざまな観点からデータを見つめること、また応用分析で重要な点は基礎分析の結果をどのような形で活用し、それをどのように新しい提案に取り込んでいくのかをグループ内で議論し尽くすことである。つまり基礎分析は浅く広く、応用分析は深く狭い範囲を対象にデータ分析することが求められるのである。こうした作業を各グループで確実に行えることができるよう、担当教員と教員を補佐する学生は、常に各グループの状況を把握するように務めている。

4. ジョイント・リサーチII

「ジョイント・リサーチII」では、経営科学系研究部会連合協議会が主催するデータ解析コンペティションに受講生全員が参加し、「ジョイント・リサーチI」で養ったデータ分析能力を駆使してより実践的なデータ分析を行っている。

データ解析コンペティションの歴史は古く1994年から毎年行われているが、その開催目的は、共通の実データを元に参加者が分析内容とその方法を競うことにある。文化情報学部がデータ解析コンペティションに参加を始めたのは2011年度からであるが、参加チームは年々増加する傾向にあり、データ解析コンペティション事務局の公式報告では2013年度の参加チームが約110チーム、参加者総数は延べ570名超まで増加している。そのため、本クラス受講生がデータ解析コンペティションに参加する際は、データ解析コンペティションの本戦に参加する前に、西日本の参加チームで関西予選を勝ち抜く必要がある。これまでの参加成績は、2011年度・2012年度・2014年度にそれぞれ関西予選で優秀賞・最優秀賞・優秀賞を受賞するなどかなり健闘はしているが、参加チームの増加・多様化に伴い、近年は関西予選ですら受賞すること自体が困難となってきている感を感じずにはおれない。

データ解析コンペティションで扱ってきたデータは、文化情報学部が参加を始めた年以降は表2のとおりであり、そのデータサイズは年々増加傾向にある[3]。そのため、「ジョイント・リサーチI」でどこまでデータベースシステムに慣れ親しみ多くのデータ分析を多角的に行ってきたかが、「ジョイント・リサーチII」での成功の秘訣となる。そのため、「ジョイント・リサーチI」と比較して、基礎分析・応用分析にかける時間も長く、さらに中間発表・最終発表前に担当教員によるデータ分析内容・分析方法のレビューも細部まで行われ、

- 基礎分析/応用分析で行っている各種データ分析が、どのような目的で行われているのか。
- 分析目的と実際のデータ分析との関係が適切か、また適切に行われているかどうか。
- 応用分析の結果が十分に考察されているかどうか。

を受講者が意識できるような講義内容となっている。

表2 データ解析コンペティション使用データ

年度	データの種類
2011	EC サイトアクセス、購買データ
2012	Web サイトアクセス、購買データ
2013	ホームスキャン、モニタアンケートデータ
2014	小売店 FSP、ID-POS データ
2015	小売店 ID-POS データ

「ジョイント・リサーチII」の具体的な授業計画は表3に示すとおりである。

- 第1回目の講義では、該当年度のデータ解析コンペティションで配布されたデータについて、毎年夏休み中に開催されるデータ解析コンペティションの発会式に参加した担当教員の研究室所属学生による説明と、配布されたデータのデータベース（通常はMicrosoft Office Access）への格納作業を行う。近年の配布データはデータサイズが大きくMicrosoft Office Accessでそのまま扱うことがサイズ的に困難であったため、データサイズを小さくするためのデータ圧縮処理やデータベースの正規化などの処理はあらかじめ担当教員側で行っておく必要があった。
- 第2～6回目の講義では、「ジョイント・リサーチI」と同様、分析対象のデータがもつ

[3] 2015年度のデータサイズはこれまででもっとも大きく、Microsoft Office Accessを用いてもデータの扱いが困難になるほどのものとなり、やむを得ず少しハードルは高くなるがOracleデータベースの使用に踏み切った。

表３「ジョイント・リサーチ II」の授業計画

実施回	内容	授業時間外の学習
1	ガイダンス（講義の目的、データ概要説明）	
2 ～ 6	基礎分析（5 回目にレビューあり）	グループ作業
7	中間発表	レポート作成
8 ～ 13	応用分析（12 回目にレビューあり）	グループ作業
14、15	最終発表	レポート作成

表４　就職実績（2012-2014）（%）

業種	2012		2013		2014	
	ビッグデータ	文化情報全体	ビッグデータ	文化情報全体	ビッグデータ	文化情報全体
メーカー	7.7	15.6	17.4	20.6	8.7	19.6
流通	0.0	16.1	13.0	14.4	4.3	10.3
金融	15.4	22.1	13.0	23.9	17.4	23.8
マスコミ・情報	0.0	16.1	8.7	16.7	0.0	20.6
教育・学習支援	7.7	3.0	0.0	1.4	0.0	2.3
サービス	30.8	11.5	21.7	15.3	34.8	12.6
公共・その他	38.5	15.6	26.1	7.7	34.8	10.8

傾向把握のための基礎分析作業を行い、その内容を中間発表で報告する準備を行う。基礎分析の期間が長いのは、データサイズや変数量が多いため、基礎分析の内容が「ジョイント・リサーチ I」よりも増加するためである。そのため場合によっては、もともと考えていたデータ分析の目的には合わないデータ分析手法を用いた基礎分析を行っている可能性もでてくる。そのようなことを防ぐため、第 5 回目の講義はレビュー日とし、各グループで行おうとしている分析の目的とその具体的なプロセスについてチェックを行っている。

- 第 8 ～ 13 回目の講義では、基礎分析で明らかになったデータの傾向を活用する手法の提案、そしてその提案の実現に向けた応用分析作業にグループで取り組み、その内容を 14、15 回目の最終発表の準備を行う。基礎分析を行う際と同様、もともと考えていたデータ分析の目的には合わないデータ分析手法を用いた応用分析行っている可能性もあるため、第 12 回目の講義はレビュー日とし、各グループで行おうとしている分析の目的とその手法についてチェックを行っている。
- データ解析コンペティションに参加する受講者は、「ジョイント・リサーチ II」の中間発表、最終発表で優秀な成績を修めた者、もしくは率先して参加を希望する者である。そうした者に対しては担当教員や担当教員の研究室の大学院生らの指導の下、発表内容のブラッシュアップのために日夜議論を戦わせている。このようなやりとりを経ることで、データ分析の本質とは何か、また、結論を得るためにどのような手段・手法をとればよいのかを経験的に学ぶことができる。

5. 本クラス受講の効果

本クラスを受講することによる効果を調査するために、本クラスを履修し単位を取得した受講生がどのような進路を歩んでいるかを過去に遡り調査してみたところ、表 4 のようになった。表中の数値は、本クラスの単位を取得した受講生全体および文化情報学部全体のうち就職先の業種を選んだ学生の割合となっている。

表 4 をみればわかるように、当該クラスの単位取得を行った学生は文化情報学部全体に比べメーカーや流通・金融といった通常の文系・理系学生が進む業種は選ばず、サービス・公共といった方面への業種に就くことが明らかに目立っている。特にサービス業へは非常に多くの受講生が進んで

おり、従来理系就職の王道でもあったメーカー系システムエンジニアを抑えている点は興味深い。また、公共・その他のうちおよそ 15 ～ 20% にあたる 3、4 名が、毎年大学院へ進学しており、これも文化情報学部全体の傾向に比べ高い値となっている。

これは近年の社会が Reich のいう 21 世紀型資本主義社会となっており、各企業が特に新しいサービスを創出するためにデータサイエンティストの能力を渇望していることが背景にあると思われる。実際、Web サービス業やコンサルタント業におけるシステムエンジニアを目指す受講生も年を重ねるごとに増えてきており、本クラスの果たす役割は非常に大きいと言える。

6. おわりに

本稿では、文化情報学部でビッグデータをテーマに扱っている「ジョイント・リサーチ I」「ジョイント・リサーチ II」において、どのような授業展開でビッグデータを題材に講義を行っているかについて述べ、またその学びが受講生の就職にどのような影響を与えているかについて考察した。その結果､本クラスを受講することで通常の文系･理系学生が選択するキャリアパスは選ばず、大学院進学やデータサイエンティストに対する需要が比較的高いサービス業（総務省，2014）に就く受講生が多いことが分かった。

今後も引き続きデータサイエンティストに対する社会の要望はますます大きくなっていくと考えられるため、より実践的なデータ分析ができる場として本クラスをより充実したものとしていきたいと考えている。

参考文献

Reich, R. B.（1992）. *The Work of Nations: Prepar-ing Ourselves for 21st Century Capitalism*, Vintage.

古賀正一（1999）.『業界を越えた技術融合による新事業の創出』『電子情報通信学会情報・システムソサエティ誌』，4（3），3.

総務省（編）（2012）.『平成 24 年度版情報通信白書』，日経印刷.

総務省（編）（2014）.『平成 26 年度版情報通信白書』，日経印刷.