１Ｃ２２

高次元スパースデータ・クラスタリングの初等的手法とその応用

○藤田裕二（政研大），白井俊行（内閣府），永松礼夫（神奈川大），宮本岩男（経産省），山本真司（河合塾）

1 はじめに

本稿では 2022 年 3 月 24 日の科学技術政策担当大臣等政務三役と総合科学技術・イノベーション会議有識者議員との会合で報告された、知識ニーズと学生の学びのギャップについての分析において用いられたクラスタリング手法について報告する。
産業界における知識ニーズと、高等教育機関が提供する学びの機会の間のマッチングについては、内閣府、経済産業省、文部科学省など複数の府省において、イノベーション創出や国際競争力などの観点から継続的な検討課題となってきた（円卓会議 2016)。「労働市場」と呼ばれる求人、求職の関係には情報の断絶、知識技能の違い、居住地域などの分断が存在するため、常に幾分かの非効率性がはさけられないものとされている。求人（欠員）と求職を平面上にプロットすると下に凸な曲線（ベバリッジ曲線）となり、求人と雇用が完全に一致することはありえない事は広く知られている（吉川 2017)。しかしながら、労働市場におけるギャップが避けられないとしても、そのギャップを放置することは事業者、労働者、社会全体にとって好ましくない。企業は激変する社会状況に対応した新たなサービスや製品を送り出すことで利潤をあげることをめざし、そのためには高度人材を雇用しようとする。高度人材は学んだことを活かすことで金銭的な成功や、経験と実績を獲得したいと願うものである。

日本の ICT に存在する問題点については長きにわたって多様な角度から分析指摘がなされてきた。ICT はハードウェア、オペレーティング・システム、ネットワーク、ユーザアプリケーション、その上に構築される新たな形態の人間社会など膨大な要素が、あるときは緊密に連結され、またあるときは緩い結合を通して思いもよらない影響を及ぼしあうことで生成展開してゆくダイナミックな領域である。ICT の問題を人材あるいは労働市場における需給の関係ととらえる指摘もあるが、もっと根源的なレベルから（林）のように、日本社会が本来的にもっているとされる変化を避ける性向にその原因をもとめた考察もある。人材だけが育っても、それを受け入れる土壌がなければ、産業として実を結ばず、社会の中で循環しつつ持続的に成長発展するプロセスが成立しないという主張である。

これらの指摘の中でも IT 人材白書（IPA2020）や産学官行動計画で言われるように、人材需給の不一致を問題視する声は、データに依拠した定量的な主張から、自分の体験や伝聞に基づく印象論まで幅広く根強く存在している。たとえば「文系 SE」などの検索フレーズを一般的な web 検索サービスで試すことで「情報処理専攻の出身でなくてもシステム・エンジニアとして活躍できる」「いやできない」といった様々な主張が、多様な媒体で多数発表されている事が確認できる。そういった、いわば一個人の印象はおくとしても、データ上で人材需給ギャップの存在を実際に確認し、もし存在するとすればその詳細を明らかにすべきである。たとえわが国の社会全体が持つ本来的性向によって ICT 分野が弱体化しているという限界の中にあったとしても、労働市場のギャップとその構造を解明し、解消することは、イノベーション創出環境の整備に資すると思われる。

内閣府総合科学技術・イノベーション会議（以下 CSTI）では 6 期基本計画のもとエビデンスに基づく政策立案とマネジメント（以下 EBPM）を推進しており政策検討上必要な分析を行うプラットフォームを e-CSTI として開発、公開している。社会人における学びのギャップは重要な分析対象の一つとして位置づけられており、対話的に操作できるシステムとして分析結果が公表されている。さらに、これをうけておこなわれた有識者会合の中でも業務の中で要求される知識と学び（学校等で履修した知

識技能）との間の不一致の存在が確認されている(有識者議員懇談会 2022 年 1 月)。また、当該有識者懇談会において ICT 領域で必要となる他業種の幅広い知識、あるいはシステム開発者とユーザという立場の違いなど ICT 領域が抱える構造的な特徴を加味したより詳細な分析を求める声があがった。
これをふまえ、(有識者議員懇談会 2022 年 3 月）では ICT 分野に就労する社会人のニーズや学生の履修状況をより詳細に分析している。具体的には、データの直接的な集計と比較以上の、データに潜在する構造的特徴を抽出し、これに基づいた分析をクラスター分析という形で実施している。しかしながら、潜在する構造的特徴は、直接にデータから読み出せないがゆえに潜在しているのであり、エビデンスに基づく政策という方向性に期待される透明性、再現可能性、客観性にとっては阻害要因ともなりかねない。

これを避けるために、本稿で紹介する手法では、利用する数学は線形代数など初等的なものに限定することで、分析手法の解釈に高度な数学的知識を必要としないよう工夫し、またデータ処理の過程でもデータに含まれない確率変数（たとえばランダム・シードなど）を避けることで、同一のデータから常に同一の結果が得られるものとして手法を構成してある。すなわち、個々の手続きや計算は古くから知られている、いわば自明なものとなっている。

2 データと手法

データは 2021 年、Web アンケートによる調査をとおして約 6700 件が取得された。調査項目は、回答者が従事している業務の種類、そこで必要となる知識からなっている。業務にはシステム・インテグレーション、アプリケーション開発、データ分析などの他、経理や総務なども含まれる。業務知識の選択肢はアルゴリズム、コンピュータ概論やデータベースなど情報処理業界で一般的なものから、ICT が重要となっている他業種の知識（創薬や機械、制御など）もカバーした 135 の選択肢から最大 5 つを選んで回答する。

以上の調査の設計により核心となるデータである業務における必要知識は各レコードが 135 の次元をもつ 2 値ベクトルである。一人が選択できる選択肢は最大 5 つなので、大部分の項は値が 0 である。このため、クラスタリングの基礎となるデータ間の類似度は大部分の場合で 0 すなわち共通要素無し、となる。すなわち、ストレートなコサイン類似度にもとづいてクラスタリングするとベクトル各要素の周辺にクラスタが局在してしまう。すなわち個別の知識ごとの分類という自明なものとなり、潜在的な構造を抽出することは困難である。

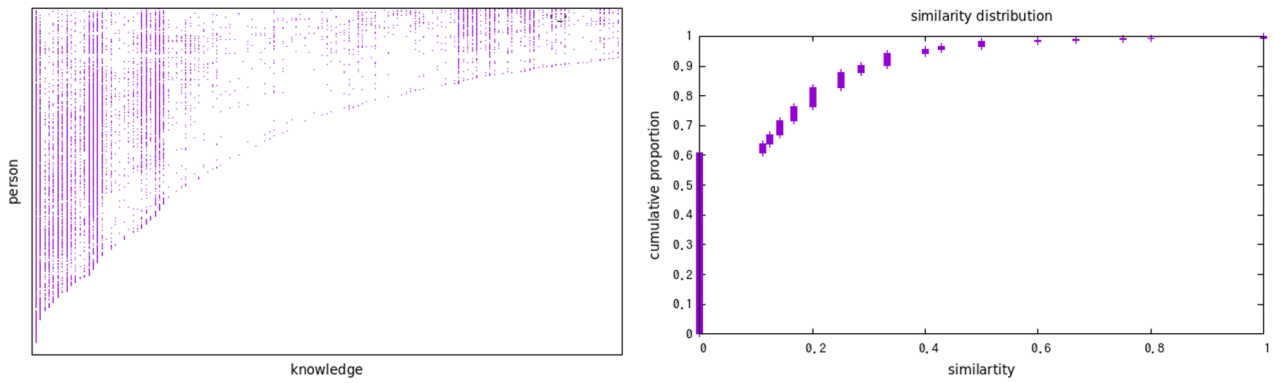

Figure 1: データ概要の可視化（a）と類似度の分布（b）

図 1a は本稿で用いたデータの概要を可視化したものであり、横方向（行ベクトル）が各回答者による個別回答である。データがスパースである事を確認しやすいように、回答中の非ゼロ数により縦方向に、また当該項が回答者に選ばれた回数で横方向にソートしてある。図 1b は、このデータから集計したコサイン類似度、すなわち x, y を回答データの非空項集合としたときの、式

$$d(x,y) = \frac{|x \cap y|}{\sqrt{|x||y|}}$$

の$d(x,y)$の値の分布であり 6 割以上で 0 となっていることが確認できる。

いま、回答者 A の[2, 30, 73, 109]という知識選択と B の[1, 9, 14, 15, 16]という選択を考える。データ上の共通点は皆無なので両者の類似度は 0 である。しかし、選択肢 1「コンピュータ概論」と選択肢 2 の「プログラミング」の間には内容上の関連があり、したがって A 氏と B 氏の接点は実際には皆無ではない。もしデータの各項の間の類似度が入手できれば、データに存在する潜在的な類似を特定できる。

そこで図 1a のデータを縦読み、すなわちデータを行列とみなして、転置（transpose）操作を施す。共通する回答者によく選ばれている知識どうしは類似しており、そうでないものは類似していないと見なすのである。知識iを特徴づけるベクトルu_iは「誰に選ばれたか」という情報を格納しており、6700 次元である。知識i,j間のコサイン類似度を保持する 135 次元の実対称行列 U を

$$U = V^t \Lambda V$$

としてスペクトル分解して得られる固有ベクトルによる正規直交基底$V = (v_1, v_2, \ldots v_{135})^t$によって各知識$W = (w_1, \ldots w_{135})$は、

$$W = V\Lambda$$

と表現される（Λ は U の固有値を保持した対角行列）。主成分分析などでは主要な少数の固有値に伴うベクトルによってデータを代表させるが、ここでの目的は知識を正規直交基底の線形和で表すことさえできれば達成されるので、情報を捨てることなく全 135 次元を用いて各知識を表現する。その上で、回答者 i を、彼らが選択した知識を表すベクトルの和$\sum w_{ik}$として表現する。

図 3 はこのようにして得られた回答者ベクトル間のコサイン類似度の分布である。図 1b に比較すると、分布が一様になっており、類似度が持つ情報量の増大が確認できる。

これらの計算はすべて、学部レベルの線形代数の教科書に掲載されている内容であり、また、広く普及している線形代数関連ライブラリを用いて、現実的な計算時間で手軽に実行可能、かつ、その実行結果も常に同一である。ライブラリの多くは自由ソフトウェアであり、ソースコードを入手して自由に改変し、再配布することができる。そのような計算環境であれば、コンピュータ内部で行われる最適化についてもすべて把握することができる。

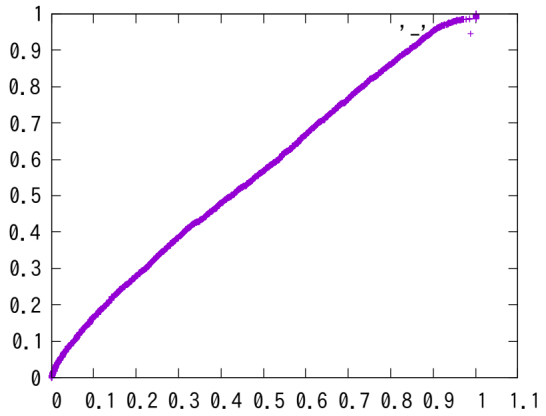

Figure 2: 潜在的知識構造に基づく回答者間類似度の分布

このようにして回答者間の潜在的な類似度が得られたとしても、その結果の品質は全データを通して一様ではない。たとえば単一の知識しか選択しなかった回答者がおよそ 1900 名存在している。そのようなデータについては、知識間の類似がそのまま個人間の類似となるが、その根拠なったデータは別の回答者の行動を含んでいる。したがって、当該回答者を特徴づけていると断言することが難しい場合も存在すると考えられる。これに対し、調査設計の上限である 5 つの知識について全て回答したデー

タはおよそ 2000 件存在する。そこで、全問回答した高品質のデータからクラスタリングを着手する。

クラスタリングの手法には様々なものがあるが、本稿で適用したのは近傍合併法 Neighbor joining である。どのような構造があるかを調査するのが目的であるから、存在するクラスタ数も未知である。また分類やクラスタリングは目的によって妥当な分割数が変化するタスクであり、一旦は分割された集合内部をよく観察すれば、更なる類似性の構造が自己相似的に一般に見出され、より詳細な分割を待っているものである。言い換えれば、客観的な、あるいは絶対的な分割の基準の存在を仮定することは一般的にいって不可能である。

この状況に対する現実的な答えは、一旦、自己相似的な類似構造を全て抽出し、そこから政策実務上妥当なレベルとなっている分割を得る手順をとることである。既述のとおり、得られている類似度はほぼ一様分布となっており、自己相似性の高いバランスのとれたバイナリ・ツリーが生成可能であり、どのようなレベルの分割にも柔軟に対応可能である。生成されたバイナリツリーから、共通した業務知識を必要としつつも、異なる業種、職種についている人材群が特定できる、14 分割となるレベルを選択した。

以上のようにして得られた分割結果$X = \{X_1, X_2, \ldots X_{14}\}$は計 2000 件のデータのクラスタリングとなっている。ここに、3 つ以上の知識を選択した回答者x_iを各クラスタの重心とx_iの類似度がもっとも高いクラスタを選んで追加してゆくことで 3900 件のデータを最終的に 14 のクラスタに分割した。下に最終的なクラスタと付随する集計結果をまとめた 2022 年 3 月 24 日開催 CSTI 有識者議員懇談会資料のページを図 4 として示す。

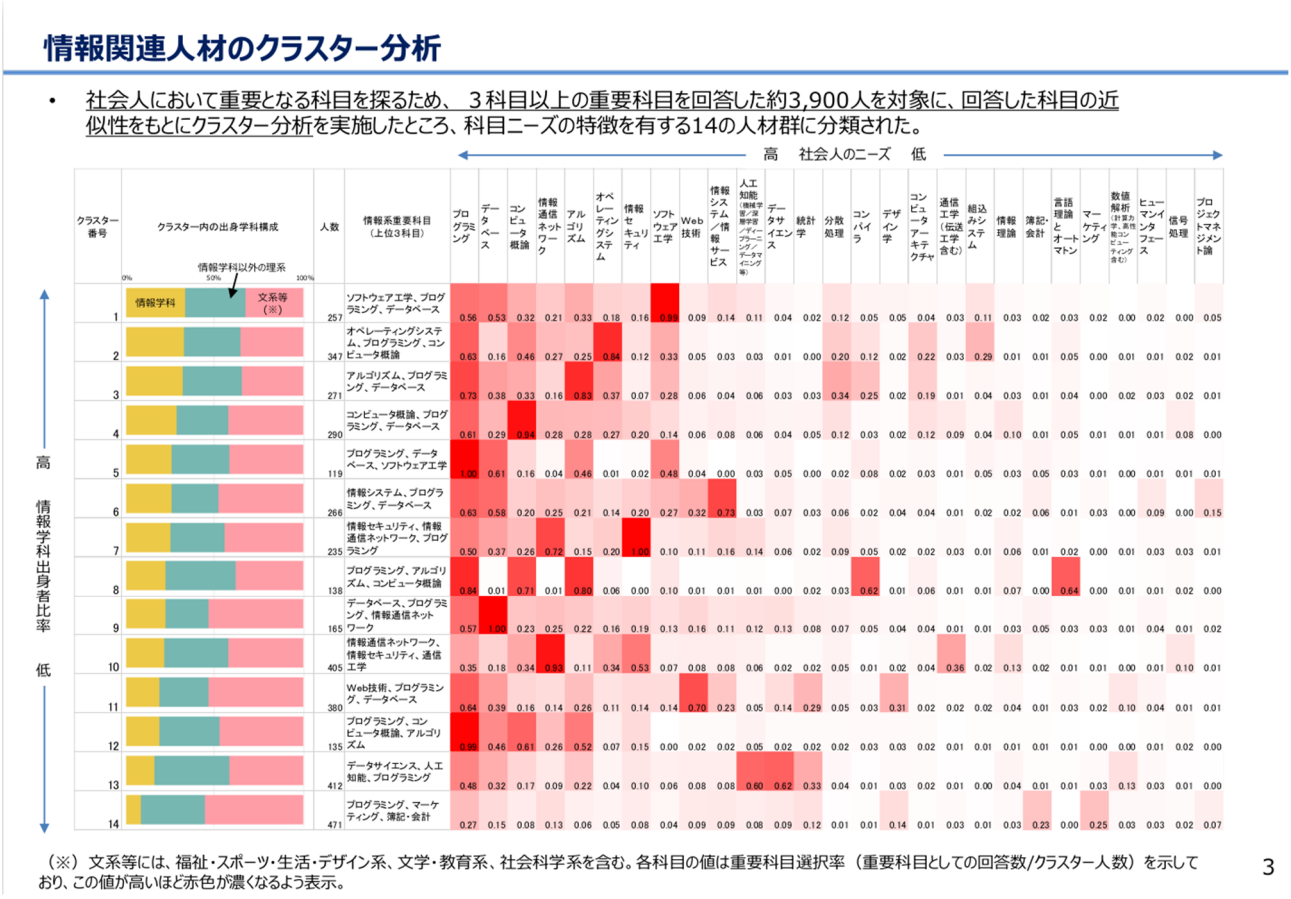

Figure 3: 有識者議員懇談会 3 月資料 P.4

3 結果と政策含意

クラスター抽出により、継続的に情報系人材の供給不足は続いているが、そのありかたが技術と産業の構造変化により変化し、新たにデータ分析業務を担える人材の不足という問題が多様な形態で発生している事が明らかになった。人材の不足は AI 開発のようなテクノロジー先行のいわば「伝統的な」ICT 領域に限らず、マーケティングや企画など従来であれば感性と経験が要求された領域もカバーできるような人材の需要が新たに発生していることが特筆される。

また、引き続き人材不足を文系出身者で補う体質が見られ、データ、AI 関連人材需要とも複合する形で人材不足が新たな局面を迎えていることが明らかになった。

> 多くの情報関連人材クラスターにおいて文系等の出身者が 3 割を超えている一方で、重要とされる科目の多くを学んでいる学生は情報学科かその他の理系に多い。プログラミングに加え、OS やアルゴリズムやソフトウェア工学といった高度な開発に必要な科目が重要な業務に携わる社会人は、情報業種に多く、情報学科出身者も比較的多いが、文系出身者も一定数存在。これらの科目を履修する学生の多くは情報学科やその他の理系学科に所属しており、文系等の学生はきわめて少ない。産業界において、本来情報分野の高度な専門知識を有する人材を獲得したいが、人材不足により、文系出身者を採用している可能性がある。
>
> データサイエンスや人工知能が重要な科目となっている社会人は、情報以外の業種や企画・営業等の事務系職種にも多く、文系出身者も多いが、これらの科目を共に履修している学生の太宗は理系である。特に文系等の学生でデータサイエンスと人工知能をともに学んでいる学生は少ない。数学の知識が前提となっており、文系等の学生には履修が困難な可能性や、そもそも履修の選択肢が提供されていない可能性がある。（2022 年 3 月有識者議員懇談会資料 p.12）

今後は階層的クラスタリング手法に近傍接合法以降に開発された手法を取り入れて実効速度や頑健性を改善するとともに、他のスパースデータ処理手法との関連を整理してまとめたいと願っている。

参考文献

（１）理工系人材育成に関する産学官円卓会議 理工系人材育成に関する産学官行動計画 https://www.meti.go.jp/policy/innovation_corp/entaku/keikaku.html　2016 年

（２）独立行政法人情報処理推進機構社会基盤センター「IT 人材白書 2020」2020 年

（３）総合科学技術・イノベーション会議有識者議員懇談会
https://www8.cao.go.jp/cstp/gaiyo/yusikisha/20220113.html
人材育成に係る産業界ニーズの分析結果について　2022 年 1 月

（４）総合科学技術・イノベーション会議有識者議員懇談会
https://www8.cao.go.jp/cstp/gaiyo/yusikisha/20220324.html
情報関連人材に関する調査結果について
～クラスター分析による社会人の知識ニーズと学生の学びのギャップの見える化の試み～　2022 年 3 月

（５）吉川洋「マクロ経済学 4 版」2017 年

（６）林晋「情報産業-日本の IT はなぜ弱いか」　イノベーション政策の科学 9 章 2015 年