2F03

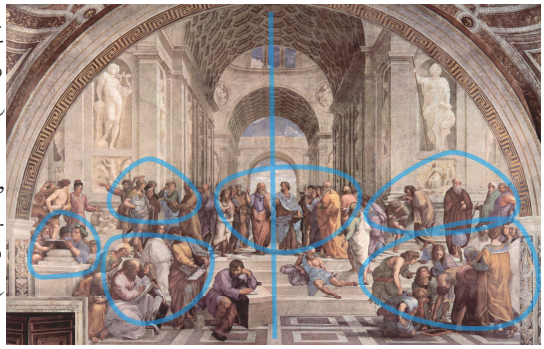
イノベーション理論の基盤としての知識構造の可視化

○藤田裕二（株式会社ターンストーンリサーチ），川口盛ノ介（株式会社盛之助），山口栄一（京都大学大学院）

要旨
知識や学問の全体構造を把握することは，常に重要である．学問や知識に限らないが，対象となる領域全体の見通しを持たないままに，最適な行動を偶然によらずとる事は不可能だからである．まして，そのような認識が視覚情報として得られるのであれば，その有用性はほぼ自明といえる．このように，自明な有用性をもった営みの例に漏れず，学問知識全般を構造化する試みには広範かつ長い歴史がある．この論文では，知識構造の可視化の歴史を振り返ることでその要件を整理し，その上でイノベーション理論という観点から，その基礎を与えるような知識構造の可視化を試み，また応用と展望を検討する．

1. 知識構造可視化の歴史と，可視化の要件

学問知識全般を構造化する試みには広範かつ長い歴史がある．西洋ではローマ時代に学問全体が 7 つの自由学芸として哲学を頂点として構造化された．この図式は以後 1000 年にわたって広く使われ，ルネサンス期にはラファエロの「アテナイの学堂」のようなすぐれた絵画芸術の画題ともなった．
ラファエロの「アテナイの学堂」は，そこで発揮された技量，画題への理解，様式と表現の斬新さと適切さなど，芸術作品として頂点を極めた存在であるのみならず，本章のテーマである知識構造の可視化としてもまた際立った存在である．



1. 人物は自由学芸，あるいはその中の主要な学説に対応し，人物の持ち物や服装，またモデルとなった人物(当時の有名人が多い)から対応を読み取れる．
2. 学問間の類縁関係が表現されている．すなわち，画面むかって左半分は詩や弁論などの感情に訴える学芸，右には自然科学あるいはその起源となった学芸が配置され，そのなかで，人々の集団として近縁な学問が表現されている．
3. 人物間のやりとりとして それらの学問間の関係性も表現されている．

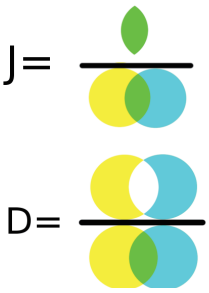
これらの項目は，適切な知識構造の視覚化が充足すべき要件として現代においても通用する．この画が壁に描かれた教皇の執務室では，当時の学問構造全体が直ちに把握可能であった．

2. アカデミックランドスケープの構築

学術振興機構採択研究開発プロジェクト「未来産業創造にむかうイノベーション戦略の研究」においても，プロジェクトの初期段階に議論と考察の出発点として，知識構造全体を俯瞰できる図，アカデミックランドスケープの必要性が痛感された．しかし現代において，ラファエロと同様のやり方で知識構造を把握し，これを視覚化することは，たとえ彼のような才能と環境に恵まれたとしても，必要な知識の量やその習得の困難さを勘案すると不可能といわざるをえない．しかし，現代には当時利用できなかったデータや分析手法そして情報処理デバイスがある．以下に，その結果とその構築手法を紹介する．

2.0 距離の概念

学問間の隔たりは，学問間の共通部分の密度で評価する．すなわち，学問 A と学問 B の記事，論文，書籍の集合を A, B とすると J(A,B)=A&B/AorB の値が小



さいほど A と B は隔たっているものとする. J は一般に Jaccard 指数として知られている関数であり, J は A と B が一致するときに最大値 1, 共通部分が存在しないとき 0 になるので, D=1-J を隔たりの尺度にする. A, B それぞれのサイズは, 公開されており誰でも利用できる web サービスである Google Scholar を利用する. これを選択した理由は, web 記事すべてを網羅する Google 本家の検索では, ノイズが多すぎるためである. 一方, Google Scholar は学術記事を主要な検索対象としており, 我々の目的により合致する. 活用するデータは, 検索結果ではなく検索ヒット件数である. A, B, A&B の三つのヒット件数がわかればこの距離は, D(A, B)= (A+B-2A&B)/(A+B-A&B) として計算可能である. D はしばしば Jaccard distance とも呼ばれる.
Jaccard distance を採用する理由はいくつかあるが, 一つは距離空間を定義する metric condition を満たすことが知られているためである. 通常, ユークリッド空間に視覚表現を構成する場合が多いので, この性質は自然な表現を得る上で望ましい. もうひとつは, 多くのデータベースで論理積検索(and 検索)が実行可能であり, その処理でも様々な最適化が利用可能な事である. したがって, データ取得という点でも容易に実装可能であり, かつ, 効率的である.

2.1 学問の選択

Wikipedia(ja)学問の一覧
http://ja.wikipedia.org/wiki/%E5%AD%A6%E5%95%8F%E3%81%AE%E4%B8%80%E8%A6%A7
にもとづいて, google scholar サービスの検索結果を勘案し, 39 の学問を選択した. 選択にあたっては人文科学, 社会科学, 自然科学を包摂し, かつ, 基礎科学と応用科学を適宜配分し,

哲学, 言語学, 心理学, 文学, 考古学, 地理学, 人類学, 政治学, 経営学, 法学, 経済学, 商学, 社会学, 物理学, 化学, 生化学, 生物学, 地学, 生命科学, 地球科学, 数学, 医学, 歯学, 薬学, 解剖学, 生理学, 栄養学, 看護学, 農学, 機械工学, 情報工学, 化学工学, 生物工学, 電気工学, 電子工学, 教育学, 情報学, 環境学, 家政学

について 2.0 の距離データが計算可能となるよう検索ヒット件数を約 1600 件収集した.

2.2 ユークリッド空間への配置

距離データは Classical Multidimensional Scaling (以下 CMDS)でユークリッド平面に布置する. CMDS は W. S. Torgerson. の発案によるアルゴリズムで, 対象間の相互距離からなる正方行列を内積に変換した後スペクトル分解し, そのうちもっとも大きな n 個の固有値に対応する固有ベクトルを n 次元ユークリッド空間における布置として用いる. すなわち, 39 次元データの 2 次元への次元圧縮でもある.
2.1 で収集した距離データの分布は, 距離の値を x, また確率密度 p(x)としたとき, 巾分布 $p(x)= x^{-2.3}$ にしたがう事がわかった. 一方スペクトル分解による近似は, 二乗誤差を最小とする解が得られる事はよく知られている. したがって, 二次の積率が収束しない分布のデータに CMDS をそのまま適用することは適切ではない.
そこで, データの即値を用いる代わりに順位を用いて, 一様分布に変換したものをスペクトル分解して座標を計算したものが, 図 1 である.

値を順位に変換すると，metric condition は成立しなくなるが，Jaccard 距離からそのまま座標を算出すると，一部の学問が一ヶ所に集結し，残りがまばらにランダムに散布された図となる．我々が必要としているのは，理論的な背景はあるが見づらい図ではなく，根拠と再現性をもった，議論の土台たりうる自然な視覚表現であるから，一様分布からの座標を選択した．1 章であげた要件のうち，最初の二つは達成することができている．残る学問間の関係性については今後の課題としたい．

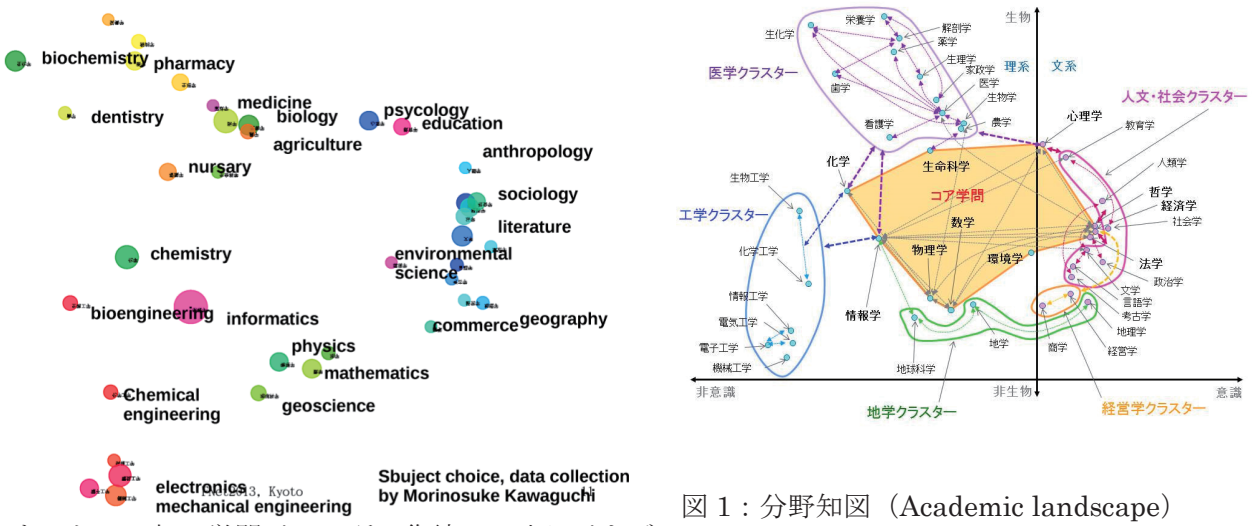


図 1：分野知図（Academic landscape）

3. Landscape から Scene へ

分野知図は知識空間全体の可視化として俯瞰を与えるものであるから，その上に技術，個人，企業などを適切に配置できれば，特定の分野に限定されない幅広い見通しを獲得できる．また客観的存在である風景に，より人間的な意味づけを与えることにもなるので，そのような活用は自然な発想である．いま，簡単のために学問が二つしか無い場合を考えて，ある対象 x と学問 A，B との結びつきの強さを a，b という数値で評価できると仮定すると，x は線分 A，B を b:a に内分する位置に配置するのが適当である． すなわち $x = aA + bB$（ただし $a, b > 0$ かつ $a + b = 1$）である．一般に， i 番目の学問との結びつきが a_i であるとすると，対象 x の座標は対応する各学問の座標 A_i として $x = \sum a_i A_i$（ただし $\sum a_i = 1$ かつ $a_i > 0$ ）である．

すなわち，結びつきの値からなるベクトル $a = (a_0 \ldots a_i \ldots), a_i > 0$ としたとき，対象の座標は固有ベクトルと a の内積である．すなわちこの座標は a から当該空間への射影である．

対象と各学問との結びつきは，2 章と同様に Jaccard index を利用する．

左図は，このようにして分野知図上にプロットした，アメリカ政府のイノベーションプログラムで成果をあげた約 5700 名の科学者あるいは技術者である．5700 名について，39 の学問との “and” 検索を行って，得られ

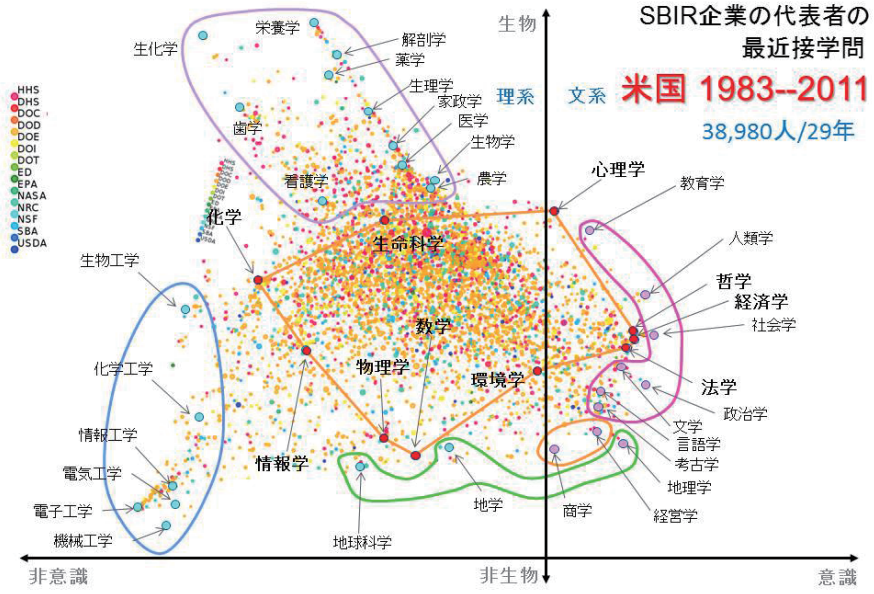


図 2：米国 SBIR 被採択企業の代表者の隣接学問

た Jaccard 指数をもとに布置を算出した. award獲得数の平方根に比例してプロットの半径を変化させ, もっとも多くの award を出した政府機関の色で塗り分けてある.

4. 結論と展望

MDS と Jaccard 指数は, 古くから利用されてきた比較的有名な手法であるが, 適用対象によっては今も十分に有用である. またデータも各種商用データベース以外に, 使い方によってはインフラ的に普及した一般的な web サービスから収集可能である. このように, 技術的には自明であっても, その組み合わせには膨大なバリエーションがあり, そこからしばしば思わぬ有用な結論を見出す事ができる場合が, まだまだあると感じている.
また, 必ずしもすべてが技術的に自明と言いきれない部分もあり, 機械学習における非線形カーネル関数と, 2 章で述べた一様分布化処理の関連なども今後調査してゆきたい.
ここで用いた手法は基本的に連続性があり, locality sensitive である. 時系列で連続的に変化するデータに適用すれば, 動きのある視覚表現が得られると考えられるので, 描画対象どうしの関連性も, これによって表現可能になるのではないか, という見通しを持っている.

謝辞
本研究は、2011 年から一貫して、科学技術振興機構(JST) 社会技術研究開発センター(RISTEX) の研究プログラム「科学技術イノベーション政策の科学」により研究助成を受けている。

参考文献

MICHAEL LEVANDOWSKY› DAVID WINTER, “**Distance between Sets”, Nature 234, 34-35**

Torgerson, W. S., “Multidimensonal Scaling: I Theory and Method”, Psychometrika vol. 17, No. 4

山口栄一**,** 藤田裕二 第 **29** 回研究・技術計画学会年次学術大会**, 2014** 年 **10** 月 **26** 日