

俯瞰解析を用いた魅力あるストーリーの抽出

データサイエンティストにむけて

Using panoramic analytics to extract insightful stories from huge volumes of text data



VALUENEX 株式会社 代表取締役社長 CEO

中村 達生

1991年、早稲田大学大学院理工学研究科を修了後、三菱総合研究所入社。1994年から1998年まで東京大学工学部助手。三菱総研復職後、2005年に工学博士取得。2006年に当社設立、代表取締役社長 CEO に就任（現任）。2014年2月米国シリコンバレーに当社現地法人を設立、CEO に就任（現任）。2018年、当社東証マザーズにIPO。現在、早稲田大学大学院理工学術院非常勤講師も兼務。2018年度特許情報普及活動功労者受章。2019年スタンフォード大学より Japan-US Innovation Award の Innovation Showcase 企業に選定。

✉ customer@valuenex.com TEL 03-6902-9833

要約

本報は情報解析により秀逸なストーリーを抽出する楽しさをお伝えすることを目的としている。

はじめに、大規模な非定型テキストデータを解析するときに留意すべき点について言及する。続いて、俯瞰解析手法の事例を用いて、特許文献を用いた解析により経営戦略等の多様な用途に利用可能であること、企業の動向をマクロ的視点から把握すること、SNS やメールから政治・政策の一端を描画することを示す。

1 はじめに

筆者は大学での研究・教育と、ビジネスの世界を交互に経験し、現在は、ビジネスに軸足を置きながら、大学にて非常勤講師を行なっている立場である。大学では、世の中の様々な事象を特許文献等の大量の文献を効果的に可視化して解析する方法とその応用事例について授業を行なっている。気がつくと、このスタイルの授業を開始してから13年もの年月が経過していたが、社会の荒波にまだ毒されていない学生が斬新な視点で作成する仮説の発表は、毎年の楽しみの一つとなっている。たとえば、ZOZO スーツの技術的視点による考察、任天堂の次のゲームの予測、トイレ休憩時の体調管理、トレーニングにおける生体データの利用、鉄道改札システムの変遷、あらたな広告媒体、ドローンの応用領域探索など、

学生が身近に感じているサービスや製品について、素直に疑問を投げかけて、関連する技術の地図（俯瞰図）を作成し、仮説の検証を行っている。そこには、彼らにとって新しい発見がたくさん存在している。あらたな疑問から仮説を修正し、さらなる解析にとりかかることを行うのである。本来、解析とは仮説設定、検証、発見、さらなる仮説設定の繰り返しであり、解析をするためのツールや方法論は、そのプロセスを支援する道具である。ビッグデータの解析ブームが起きて久しいが、いま、もっとも求められているのは仮説設定能力の向上だろう。本報では方法論に頼りきるのではなく、俯瞰解析の可能性と解析事例を示し、各事例の解析を通じて抽出されているストーリーの一端を紹介するとともに、解析することの楽しさを知っていただくことを目的としている。

2 大規模な非定型テキストデータ解析

2.1 解析の心得

大規模な情報を対象に解析を行うとき、客観性と俯瞰性を意識することが大切である。俯瞰性とは目的に照らして十分な範囲と期間の情報を用いることであり、客観性とは恣意性の排除である。人の知見、判断を極力排除し、欠けた情報をつなぎあわせて、全体像を把握することが肝要だ。また、手法に溺れず、データを素直にみて、データは不完全であることを認識する。完璧なデータクリーニングは不可能であるし、データコレクターになっ

てはいけない。残念ながら、すべての課題に適用できる王道の手法は存在しない。日本の知財業界を生業とする領域において、2017年よりIPランドスケープという言葉が流行しているが、これはコンセプトであり手法ではない。セミナー等のトレーニング教室は存在しても、実践的な手法は学べない。データから情報に昇華するには、適切な範囲のデータを用い、精度の高い解析手法を用いる必要がある。情報から知識を引き出すには、欠けている情報の存在を理解することが求められる。さながら、宇宙空間の80%以上を占めるダークマターを理解するに等しい。解析プロセスを複数の段階に分けると、人が介在しなければならないのは、最初の仮説設定と最後の結論導出の段階である。現在のどのようなテクノロジーをもってしても、この二つの段階は、自動化ができない。解析ツールのボタンを押すとグラフやテーブルは出て来るが、判断を伴う結論を期待することは、残念ながら2030年以降に実現するかもしれない汎用人工知能が出てくるまでは困難だろう。この人智にかかわる部分が秀逸なストーリーを抽出する上で欠かせないが、経験を積むと生みの喜びを味わえるようになる。

2.2 特許文献を用いた解析は戦略立案に有用

特許文献は網羅性、情報の品質、経年変化の取得に優れているだけでなく、付帯情報を利用して、人や組織間のつながりから今後の潮流を読み込むことに適した情報源の一つである。権利書としての側面だけではなく、優れた技術書としての側面に着目して、将来に向けた解析に利用できる。マーケット情報、金融情報等を融合することにより、特許調査、M&A デューデリジェンス、株式調査、投資戦略、研究開発戦略、競合分析、ヘッドハンティングなどの様々な分析目的に資することが可能で

ある(図1)。

2.3 俯瞰性と客観性を実現するための精度はどの程度必要か

粗い表現だが、10,000件の特許文献を分析して区別するには、最低でも10,000の次元が必要だ。単純に考えて10,000の発明があると考えるとわかりやすい。ところが通常のワークステーションの構成では256次元が最大次元となる。この程度の次元では精度が出ない。多くの解析ツールでは、いかに次元を落として短時間で計算をするかに力が注がれ、精度を担保する方法はあまり注目をされてこなかった経緯がある。このため、自社の複数の特許文献を含めた分野全体の解析をすると、いつのまにか、いくつかの自社の特許文献が行方不明になるという事件が発生する。なぜなら、低次元の要素数ではすべての文献にタグ付けすることができないからである。分析の目的に照らして、必要な結果を生み出すためには、量と時間と精度のバランスがとても重い意味を持つ。

2.4 目的に適合した可視化手法の探索

大量のドキュメントは何かしらの手法で可視化する必要がある。可視化手法はMDS (Multi-Dimensional Scaling)¹ や SOM (Self Organizing Method)²、またはこれらの派生種が主流であり、その主な目的は既存の分類にふりわけることであるため、既存分類によらない

1 千野直仁, 佐部利真吾, 岡田謙介, 「非対称 MDS の理論と応用」現代数学社

2 Gintautas Dzemyda, Olga Kurasova, Julius Zilinskas, "Multidimensional Data Visualization", 2013



図1 特許文献の俯瞰解析にてできること

新しいクラスタリングを生成することは得意ではない。一般的な MDS は次元を縮退させて 2 次元に表現する方法であるため、最終的に遠い三つの関係は正確に距離で表現できるが、近い関係は誤差の中に潰れてしまうため、いくらズームインしても意味がない。ヒートマップのように色をつけて、分かりづらくしていても、中心から放射状に広がる三本の線に沿って、ドットが描かれていることが特徴である。SOM はあらかじめ作成したフレームの上に、プロットしてゆく描き方となり、限られた少ない次元でしか表現できないため、前出のように投入したはずの文献が行方不明になるという事象が当たり前のように発生する。おそらく、データベースサーブスに付随したこの手のツールを使ったことがある人は、思い当たることがあるだろう。多少の速度が犠牲になっても、意味ある結果を出すための可視化手法を利用することが必須である。精度の良し悪しは、実際に利用して自分の馴染みのある分野について可視化し、ひとつひとつの関係が納得いくものになっているかを確認することで証明できる。次節では、精度を重視した可視化手法を用いた可視化と解析事例を紹介する。

3 解析事例にみる我が国産業の優位性と衰退

我が国の代表的な企業の研究開発動向を俯瞰して戦略の変遷について考察を行った。

3.1 事業転換の成功ストーリーは研究開発段階からの目利き力

デジタルカメラが市民権を得て、市場に爆発的に広がった 2003 年頃、フィルムメーカー各社は岐路に立たされていたと、誰もが容易に思い描ける。研究開発活動の推移を可視化すると、富士フィルム、Kodak、Konica の主要三社は三様の事業転換を行なっていたことが見えてくる。富士フィルムはフィルムの基幹技術の一つであるエマルジョン（乳化剤）を使って、化粧品事業への華麗なる転換を行い、成功ストーリーとして世に広く認知されている。ところが、富士フィルムの 1980 年初期の特許文献の中に、すでに、化粧品への応用が言及されており、研究開発重心も毎年少しずつ動いており（図 2）、決して 2003 年になって急激に研究開発のジャンプが起きているのではない。応用領域も含めた視点で日頃から研究開発を行なっている実績が、事業継続性と事業転換の両輪を実現する強さに現れていると言えよう。一方の Kodak は研究開発の重心がフィルムの領域から移動せずに、そのまま事業が消滅している。ただし、Kodak の知財権の一部は米国の大手企業群に引き継がれており、研究開発の投資収支の視点からは決して無駄にはならず済んでいる（図 3）。Konica はミノルタと合併するとともに、フィルム事業から撤退しており、研究開発の重心も直角に方向転換をして、事業体としての生き残りに成功している。このように成熟分

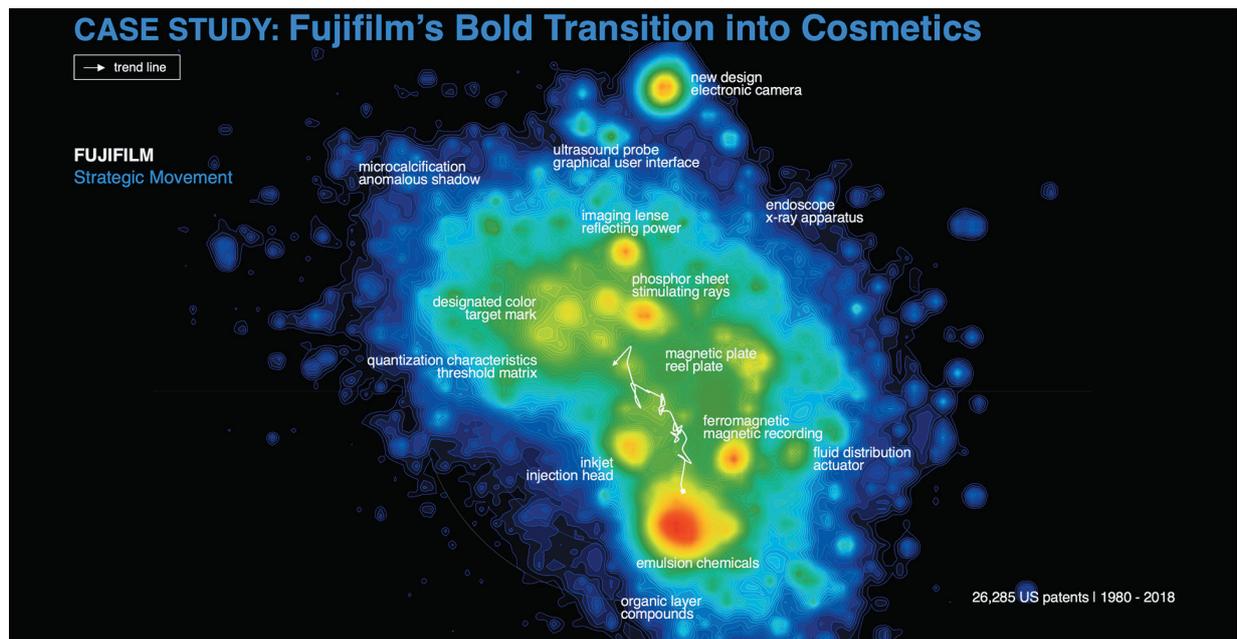


図 2 富士フィルムの技術俯瞰と重心の動き

野の末期における三者三様のストーリーが特許文献をベースに用いて引き出すことが可能である。

3.2 成熟産業の将来ビジョン策定の必要性

鉱山事業のようなインフラ系企業などの安定した成熟産業では、研究開発の重心が過去 20 年以上も動かない複数の企業事例が存在する。今後を見通した時に、埋蔵量はいまだに数十年単位で存在しているため、現在の

事業は当分の間は安泰であるように思える。しかし、市場側のニーズ、環境の変化にともない、見えない変化が迫っている。再生可能エネルギーによる代替の進行、技術革新による代替材料の低価格化と構造材への適用の拡大は、巨大市場だが、利益率が数パーセントという産業へのインパクトは大きい。巨艦の成熟産業であるがゆえに変革に時間を有することが想定されるため、考えられる中長期的なビジョンを早期の段階から策定する必要が

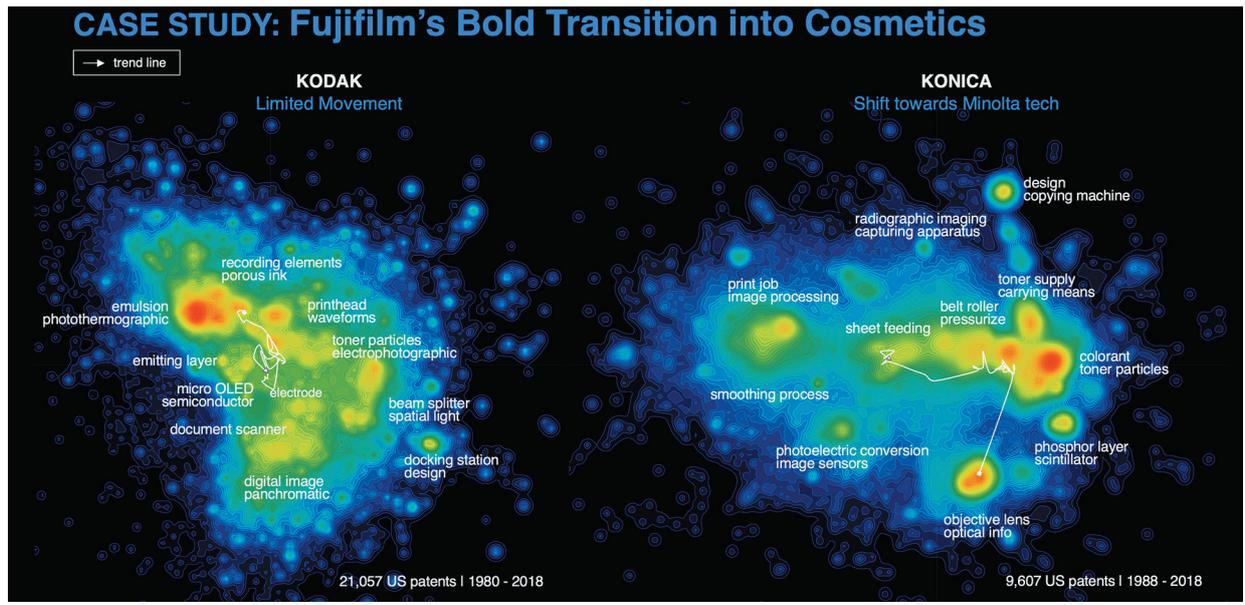


図 3 KODAK と KONICA の技術俯瞰

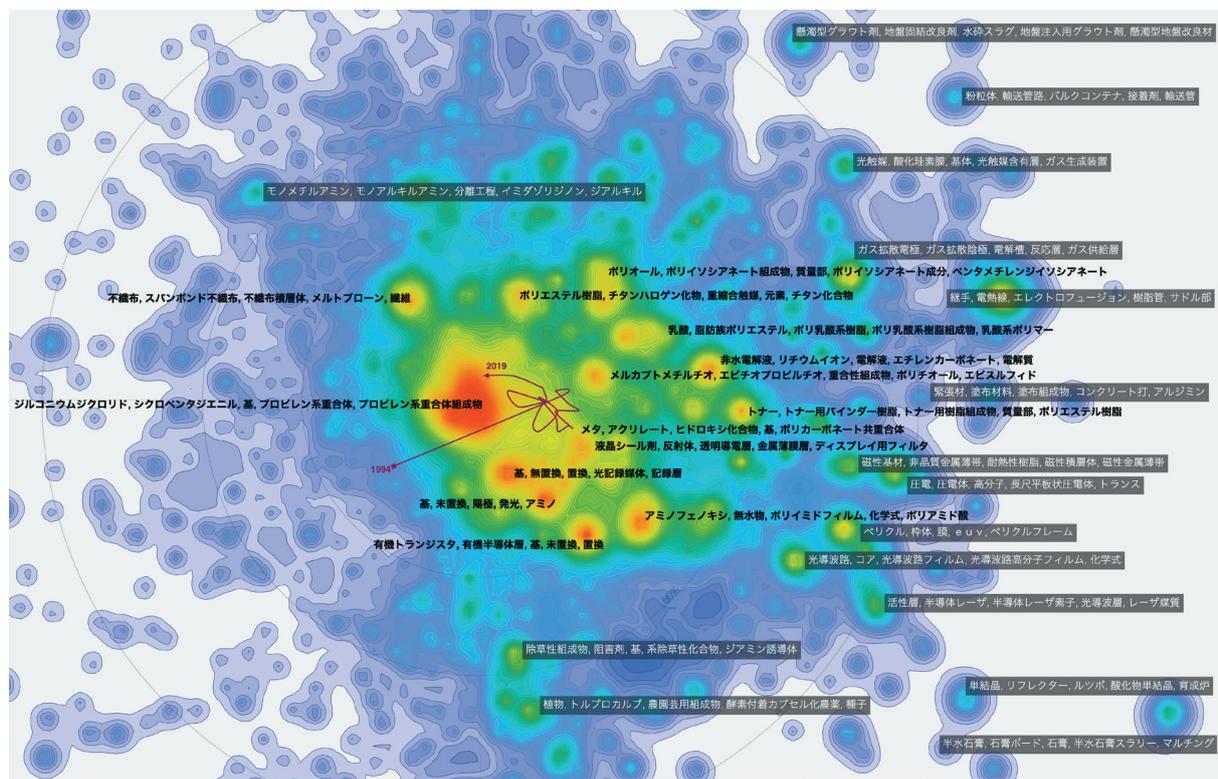


図 4 約 24 年も研究開発の重心が動いていない化学メーカー

あり、社会動態、生活の動向を踏まえた研究開発戦略の策定が望まれている。

3.3 強みであった動かない研究開発の重心が弱みに転じる

研究開発の重心に動きがないことは、社会インフラ産業に限った現象ではない。キヤノンも過去 10 年以上にわたり研究開発の重心が動いていない企業の一つである。キヤノンは複写機、光学機器、半導体露光装置、インクジェットプリンターなど独立した複数の強い事業が存在し、それぞれに紐づく研究開発活動がある種のバランスを維持したまま行われてきた。キヤノンの特許文献を用いた俯瞰図を作成すると、複数の密度の濃い領域と各年の重心がほぼ一箇所に止まっていることが確認できる(図 5)。売り上げ推移は各事業それぞれ異なるはずであるが、特許の出願件数には影響していないように見える。特定の事業が苦しいときにも、その事業部門における研究開発の手綱を緩めずに遂行したことの表れとみることができ、これが事業を盛り返す機会の増加につながっているのではないかと推察できる。研究開発の重心が動かないことは、1990 年代、2000 年代においてキヤノンの市場における強みにつながっていた。ところ

が、カメラのスマホ代替、印刷需要の減少等、市場環境が変化する現在は、画一的な分野での研究開発活動が弱みに転じてきている可能性がある。このような複数の事業部門を擁する大企業の場合は、事業間、もしくは視点を広げて、事業の周辺に存在する萌芽領域を探索することを日頃から実施していることが重要である。

3.4 プラットフォーマーであることに気づかない企業

GAFGA³によるクラウドサービスを基本としたプラットフォーム事業が世界を席卷している。メール、データストレージ、ホームスピーカー、配車アプリなど、プラットフォームを前提としたサービスは基本的に無料、または低価格にて提供されている。したがって、これらの他社が構築したプラットフォームで稼働するサービスやデバイスの販売は基本的に薄利となる。事業として成功するには、自らがプラットフォームになるのが正しいことは、胴元が常勝する道理と同じである。そのためには、巨人のプラットフォームが手がけていない領域を見つけ出すことから始めなければならない。GAFGA が出願

3 Google, Apple, Facebook, Amazon の総称

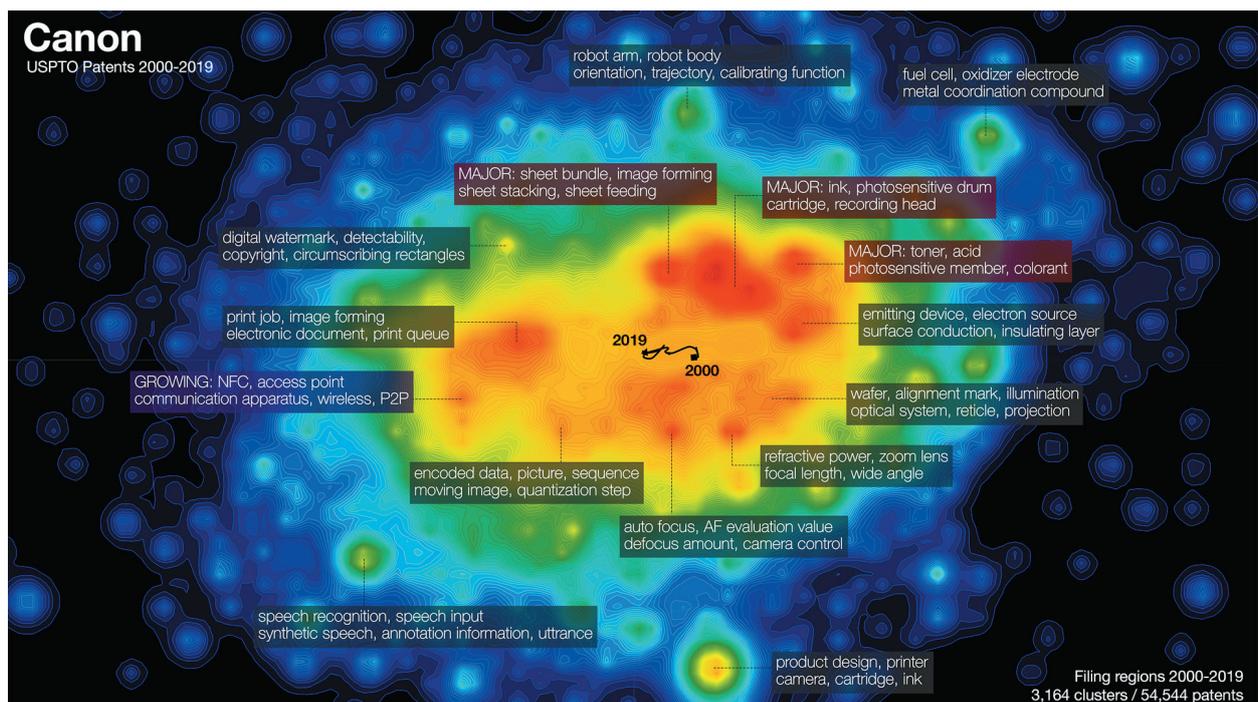


図 5 安定しているキャノンの研究開発重心

した特許文献を俯瞰図に表すと、彼らが注力している研究開発領域間に、比較的大きな隙間（White Space）が存在していることがわかる（図6）。この隙間は彼らの何かしらの理由で出願件数が少なく、すなわち、研究開発活動が手薄な領域といえよう。この隙間領域に強みの技術を保有する企業は、GAFA と対峙したビジネスを起こせる可能性を秘めている。このような俯瞰図を作成するまで、自らがすでに産業のプラットフォーマーであることに気づいていない企業も存在する。前節で取り上げた重心が長期にわたり動かない企業は、その分野の基礎的な材料やデバイスの提供を行い、高い製品シェアを実現しているが故に研究開発の重心が動かない技術深耕型の企業である。コア技術を中心としたサービス化の展開や周辺事業へのライセンス等は、すなわち、プラットフォーム事業の一種と呼ぶことができる。

4 政治・政策の俯瞰

情報解析は政治・政策の分野へも浸透しつつある。政治家の公式文書での発言だけでなく、個人的に SNS へ投稿された文章やメールのログデータは、数を集めることにより背景にある本音やキーパーソン、キーワード

を浮かび上がらせることができる。過去の4人のアメリカ大統領が発言した一般教書演説は数万ページにもものぼる長大な文書である。これらをつぶさに読み込み総括することは大変な労力を伴う。しかし、各元大統領のドキュメントを俯瞰図に可視化し、4人の図を並べると、各元大統領の政策の力点の違いを明確に示すことができる（図7）。一方、現職のトランプ大統領は就任前から Twitter 上でつぶやきを発信し続けており、その14,622件のツイートを用いて可視化を行なってみた（図8）。一般教書演説とは異なり、短文かつ指示代名詞が多用されているため、やや散散的な俯瞰図に仕上がっている。中心より左側に就任前の発言が多く、Fake など候補者を批判する単語を含むツイートや Fox News が特徴語として上がってきているのがわかる。図の右側には就任後の発言が多く、あらゆる内容の発言がツイートされているが、唯一ブレない発言が“America First”であることが興味深い。

トランプ大統領を相手に選挙を戦ったヒラリー・クリントンは、2015年3月に公務情報を私用メールアカウントで利用したとして問題になり、その後の選挙戦で窮地に立たされた。その話題となった一連のメールのデータはインターネット上で公開されており、内容や

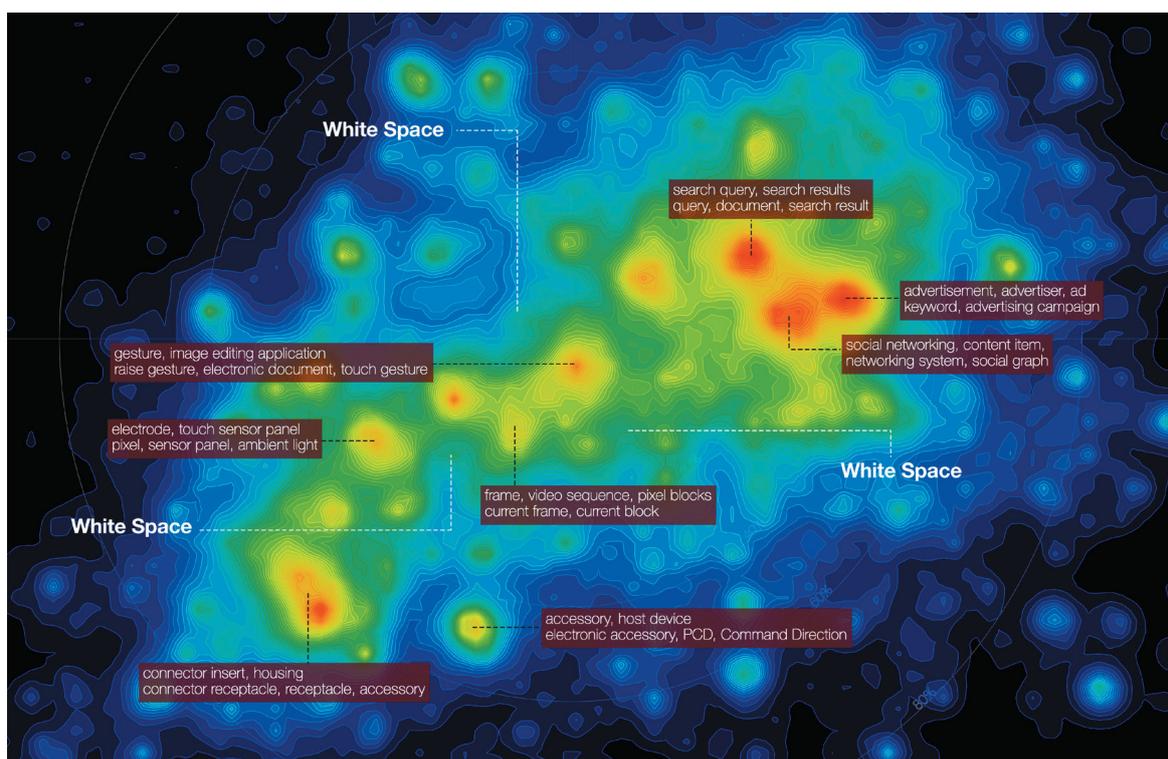


図6 GAFAの開発領域の隙間

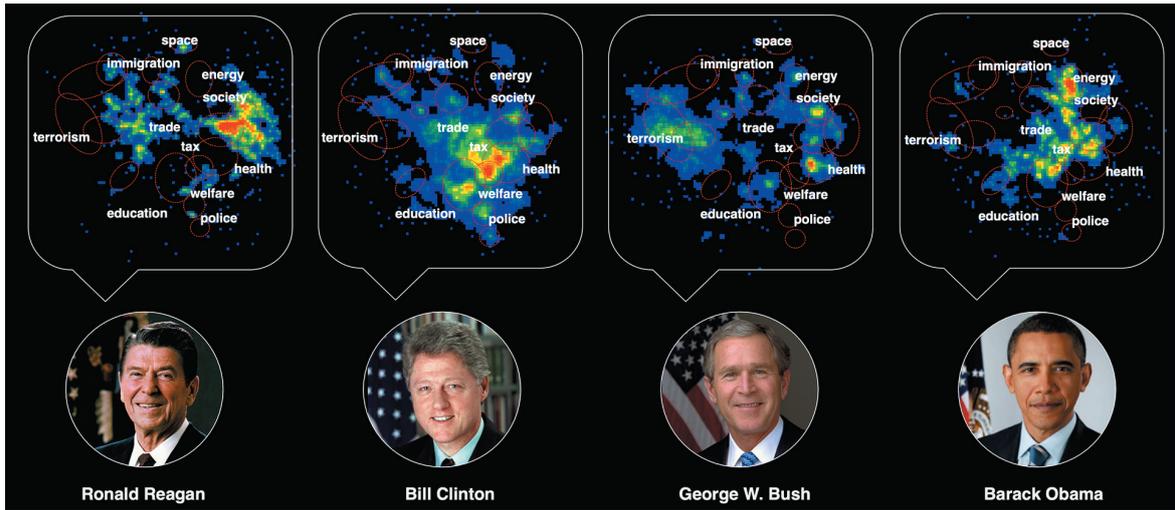


図 7 過去のアメリカ大統領の一般教書演説の俯瞰

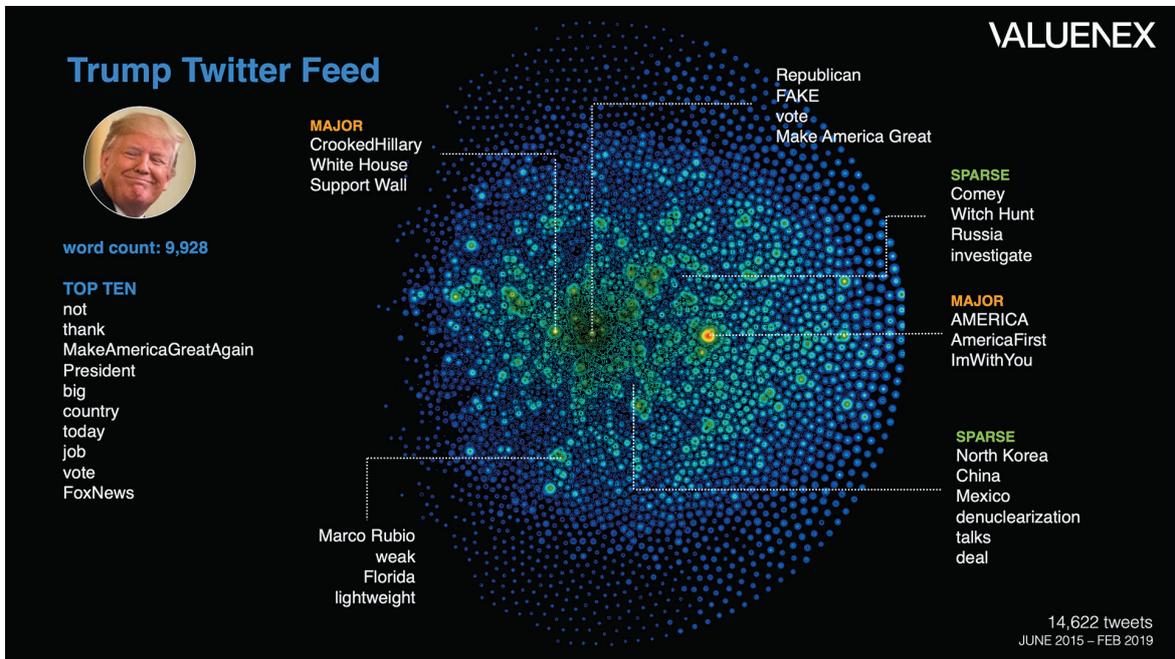


図 8 トランプ大統領の Twitter の俯瞰図

ヘッダー情報を用いることにより、ヒラリー・クリントンを取りまく人的なつながりの関係図を可視化することができる(図 9)。選挙に関する単語を多用するグループがおそらく選挙対策チームであり、その中心に位置する人物がリーダーなのだろうと推測をつけることができる。この解析の延長上には当選予測への適用が期待されるが、現段階では複雑な要素の組み合わせや、そもそものデータ取得の制約があり、実現にはまだ時間を要するだろう。

5 結言

特許文献、論文、SNS 等の公開情報を活用することにより、産業構造の変化、経営戦略、政治・政策の分析に資する情報が得られる。分析の目的によらず、課題の設定、適切な情報ソースの選定、漏れのない必要十分な量の情報の取得、広く俯瞰して大局を掴む、可視化・分析の結果の検証、結論の導出の多段のステップを経る必要がある。単体のステップ、または複数のステップを支援するツールやサービスが存在するが、適用可能範囲や必要な精度を十分に考慮して活用しなければならない。

