IIR

Internet
Infrastructure
Review

Sep.2020
Vol.48

定期観測レポート

ブロードバンドトラフィックレポート ～新型コロナウイルス感染拡大の影響～

フォーカス・リサーチ(1)

5G時代のMVNOの在り方 ～VMNO構想の実現に向けた取り組み

フォーカス・リサーチ(2)

Splunkによる日本語文章解析処理

IIJ
Internet Initiative Japan

Internet Infrastructure Review

September 2020 Vol.48

エグゼクティブサマリ

日本では、固定系ブロードバンド契約者のトラフィックの実態を把握するため、総務省が主要なインターネットサービスプロバイダ、インターネットエクスチェンジ、研究者の協力を得て、トラフィックの集計・試算を行っています。この調査はIIJも参加して2004年から続けられており、インターネットの発展を記録するうえで、非常に貴重なデータの1つとなっています。この度、最新版の2020年5月の集計結果が公表[*1]されました。今年は新型コロナウイルス禍によって、インターネットのトラフィックが世界的に増大していることはメディアでも報道されており、皆さまもご存じのことでしょう。今回、集計対象となったのは、緊急事態宣言が発令されて、国民の移動が最も厳しく制限されていた時期と重なります。結果として、固定系ブロードバンド契約者の総ダウンロードトラフィックが、昨年5月は前年度比17.5%の増加であったのに対し、今回は前年度比57.4%と大幅な増加になっています。総アップロードトラフィックも48.5%の増加と大きな伸びを示しています。新型コロナウイルス禍がインターネットのトラフィックに与える影響があらためて数字として明らかにされたと共に、非常時においてインターネットが果たした役割の大きさも読み取れる貴重なデータであると受け止めています。

今回の新型コロナウイルス禍に限らず、地震や台風といった自然災害など、社会に多大な影響を与える事象が発生したとき、インターネットは大きな役割を果たしてきました。そしてこれからもインターネットが社会インフラとして期待される役割を全うできるよう、我々も努力していきたいと考えております。

「IIR」は、IIJで研究・開発している幅広い技術を紹介しており、日々のサービス運用から得られる各種データをまとめた「定期観測レポート」と、特定テーマを掘り下げた「フォーカス・リサーチ」から構成されています。

1章の「定期観測レポート」では、IIJの固定ブロードバンドとモバイルのトラフィックの分析結果を報告しています。総務省の集計をもとに新型コロナウイルス禍の影響が出ていることは先述しましたが、こちらの分析では、より詳細にその影響を見ることができました。総トラフィックは、今年に入ってから国民の行動が制限され、5月の緊急事態宣言でそれがピークに達し、6月に入って緩和されたことが明確に読み取れました。また、6月初旬の利用者当たりのトラフィックの量分布は、固定ブロードバンドが増大、モバイルが減少という結果となり、国民の行動が制限され自宅での活動が増加したことが、トラフィックからも裏付けられる結果となりました。

2章の「フォーカス・リサーチ」では、5G時代におけるMVNOの在り方として我々が提唱するVMNO構想を解説しています。昨年から世界各国で5Gのサービスが開始され、日本でも今年からMNOの本サービスが始まりました。しかし、現行の5Gサービスは既存の4Gの仕組みのもと無線区間だけ5Gの仕組みを取り入れて超高速通信を実現しているもので、NSA方式と呼ばれています。NSA方式においては、MNOとMVNOの関係は4Gの時代と変わりありません。5Gで検討されてきた多数同時接続、超低遅延を実現するには、本格的に5Gの仕組みを採用したSA方式に移行する必要があります。VMNO構想は、MVNOがSA方式のもとで5Gの特徴を活かしたサービスを提供するための考え方を提案しています。

3章の「フォーカス・リサーチ」では、我々が大規模メールサービスを運用するなかで、機械学習の技術を活用して、スパム検知の自動化、サービス運用の効率化を実現してきた取り組みを紹介しています。その取り組みではSplunkを利用していますが、SplunkのNLP（Natural Language Processing）が日本語に対応していなかったため、独自でNLPを拡張し、日本語のテキストマイニングを可能にしました。テキストマイニングのビジネス活用の事例としても参考にしていただければと思います。

IIJは、このような活動を通してインターネットの安定性を維持しながら、日々、改善・発展させていく努力を行っています。今後も企業活動のインフラとして最大限に活用いただけるよう、様々なサービスやソリューションを提供し続けてまいります。

島上 純一（しまがみ じゅんいち）
IIJ 取締役 CTO。インターネットに魅かれて、1996年9月にIIJ入社。IIJが主導したアジア域内ネットワークA-BoneやIIJのバックボーンネットワークの設計、構築に従事した後、IIJのネットワークサービスを統括。2015年よりCTOとしてネットワーク、クラウド、セキュリティなど技術全般を統括。2017年4月にテレコムサービス協会MVNO委員会の委員長に就任。

＊1　総務省、「我が国のインターネットにおけるトラヒックの集計・試算」（https://www.soumu.go.jp/menu_news/s-news/01kiban04_02000171.html）。

ブロードバンドトラフィックレポート ～新型コロナウイルス感染拡大の影響～

1.1 概要

このレポートでは、毎年IIJが運用しているブロードバンド接続サービスのトラフィックを分析して、その結果を報告しています[*1][*2][*3]。今回も、利用者の1日のトラフィック量やポート別使用量などを基に、この1年間のトラフィック傾向の変化を報告します。今回、新型コロナウイルス感染拡大の影響で、自宅でのインターネット利用が大幅に増え、それに伴いブロードバンドトラフィックも増加しました。その一方で、外出が減った分モバイルの利用は減少しています。

図-1は、IIJの固定ブロードバンドサービス及びモバイルサービス全体について、月ごとの平均トラフィック量の推移を示したグラフです。トラフィックのIN/OUTはISPから見た方向を表し、INは利用者からのアップロード、OUTは利用者へのダウンロードとなります。トラフィック量の数値は開示できないため、両サービスの1年前の2019年6月のOUTの値を1として正規化しています。

ブロードバンドサービスのトラフィックは、新型コロナウイルスの国内感染が本格的に広がり始めた3月から5月にかけて急増し、緊急事態宣言解除後の6月には少し減りました。この期間の詳細については前号の記事で報告しています[*4]。この1年のブロードバンドトラフィック量は、INは43%の増加、OUTは34%の増加となっています。1年前はそれぞれ12%と19%の増加でしたので、大幅に増加したことが分かります。逆に、モバイルサービスは、リモートワーク向けの利用が増えたものの、外出時の利用分が大幅に減ったため、全体ではこの間にトラフィックが減少しました。その後、こちらも6月には少し戻してきています。モバイルは、この1年で、INは28%の増加、OUTは7%の減少と、初めてダウンロードが減少しました。1年前はINが60%、OUTが22%の増加でした。

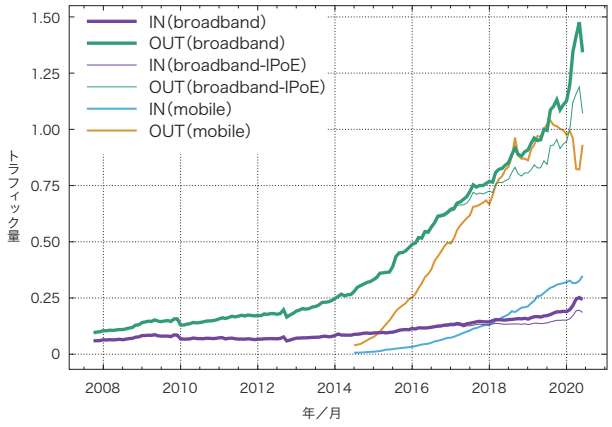


図-1 ブロードバンド及びモバイルの月間トラフィック量の推移

ブロードバンドに関しては、IPv6 IPoEのトラフィック量も含めて示しています。IIJのブロードバンドにおけるIPv6には、IPoE方式とPPPoE方式がありますが[*5]、IPoEトラフィックはインターネットマルチフィード社のtransixサービスを利用していて直接IIJの網を通らないため、以降の解析の対象にはなっていません。2020年6月時点で、IPoEのブロードバンドトラフィック量の全体に占める割合は、INで24%、OUTで20%と、昨年同月よりそれぞれ5ポイントと6ポイント増えています。特に3月以降はPPPoEの輻輳が目立つようになり、それを避けてIPoEへ移行する利用者が増えていて、IPoEの利用拡大が加速しています。

1.2 データについて

今回も前回までと同様に、ブロードバンドに関しては、個人及び法人向けのブロードバンド接続サービスについて、ファイバーとDSLによるブロードバンド顧客を収容するルータで、Sampled NetFlowにより収集した調査データを利用しています。モバイルに関しては、個人及び法人向けのモバイルサービスについて、使用量にはアクセスゲートウェイの課金用情報を、使用ポートにはサービス収容ルータでのSampled NetFlowデータを利用しています。

トラフィックは平日と休日で傾向が異なるため、1週間分のトラフィックを解析しています。今回は、2020年6月1日から6月7日の1週間分のデータを使っていて、前回解析した2019年5月27日から6月2日の1週間分と比較します。

＊1　長健二朗. ブロードバンドトラフィックレポート: トラフィック量は緩やかな伸びが継続. Internet Infrastructure Review. Vol.44. pp4-9. September 2019.
＊2　長健二朗. ブロードバンドトラフィックレポート: ダウンロードの増加率は2 年連続で減少. Internet Infrastructure Review. Vol.40. pp4-9. September 2018.
＊3　長健二朗. ブロードバンドトラフィックレポート: トラフィック増加はややペースダウン. Internet Infrastructure Review. Vol.36. pp4-9. August 2017.
＊4　長健二朗. 新型コロナウイルスのフレッツトラフィックへの影響. Internet Infrastructure Review. Vol.47. pp18-23. June 2020.
＊5　小川晃通, 久保田聡. 徹底解説v6 プラス. ラムダノート. January 2020.(https://www.jpne.co.jp/books/v6plus/)。

ブロードバンドの集計は契約ごとに行い、一方モバイルでは複数電話番号の契約があるので電話番号ごとの集計となっています。ブロードバンド各利用者の使用量は、利用者に割り当てられたIPアドレスと、観測されたIPアドレスを照合して求めています。また、NetFlowではパケットをサンプリングして統計情報を取得しています。サンプリングレートは、ルータの性能や負荷を考慮して、1/8192 ～1/16384に設定されています。観測された使用量に、サンプリングレートの逆数を掛けることで全体の使用量を推定しています。

IIJの提供するブロードバンドサービスにはファイバー接続とDSL接続がありますが、今ではファイバー接続の利用がほとんどとなっています。2020年には観測されたユーザ数の98%はファイバー利用者で、ブロードバンドトラフィック量全体の99%以上を占めています。

1.3 利用者の1日の使用量

まずは、ブロードバンド及びモバイル利用者の1日の使用量をいくつかの切り口から見ていきます。ここでの1日の利用量は各利用者の1週間分のデータの1日平均です。

前回から、利用者の1日の使用量は個人向けサービス利用者のデータのみを使っています。これは、利用形態が多様な法人向けサービスを含めると分布の歪みが大きくなってしまうため、全体の利用傾向を掴むには個人向けサービス分だけを対象にした方が、より一般性がありかつ分かりやすいと判断したからです。なお、次章のポート別使用量の解析では区別が難しいため法人向けも含めたデータを使っています。

図-2及び図-3は、ブロードバンドとモバイル利用者の1日の平均利用量の分布（確率密度関数）を示します。アップロード（IN）とダウンロード（OUT）に分け、利用者のトラフィック量をX軸に、その出現確率をY軸に示していて、2019年と2020年を比較しています。X軸はログスケールで、10KB（10^4）から100GB（10^{11}）の範囲を示しています。一部の利用者はグラフの範囲外にありますが、概ね100GB（10^{11}）までの範囲に分布しています。

ブロードバンドのINとOUTの各分布は、片対数グラフ上で正規分布となる、対数正規分布に近い形をしています。これはリニアなグラフで見ると、左端近くにピークがあり右へなだらかに減少する、いわゆるロングテールな分布です。

OUTの分布はINの分布より右にずれていて、ダウンロード量がアップロード量より、ひと桁以上大きくなっています。2019年と2020年で比較すると、INとOUT共に分布の山が右に移動しており、利用者全体のトラフィック量が増えていることが分かります。今回は前回に比べてこの増加量が大きくなっています。

右側のOUTの分布を見ると、分布のピークはここ数年間で着実に右に移動していますが、右端のヘビーユーザの使用量はあま

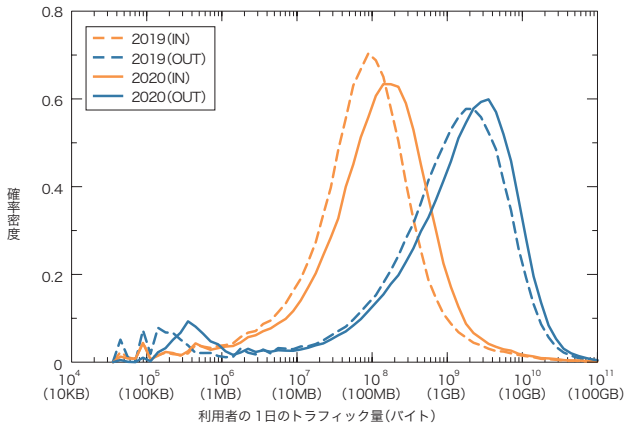


図-2 ブロードバンド利用者の1日のトラフィック量分布
2019年と2020年の比較

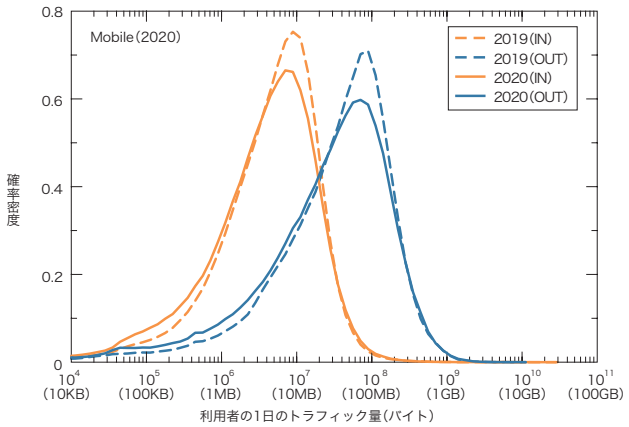


図-3 モバイル利用者の1日のトラフィック量分布
2019年と2020年の比較

り増えておらず、分布の対称性が崩れてきています。一方で、左側のINの分布は左右対称で、より対数正規分布に近い形です。

図-3のモバイルの場合、逆に分布の山が若干左に移動していて、山の高さが低くなって山の左側が持ち上がっているのが分かります。これは、利用量が多い利用者割合はあまり変わっていないが、利用量が中間ぐらいの利用者割合が減って、利用量が少ない利用者割合が増えていることを示しています。

モバイルの利用量は、ブロードバンドに比べて大幅に少なく、また、使用量に制限があるため、分布右側のヘビーユーザの割合が少なく、左右非対称な形になります。極端なヘビーユーザも存在しません。外出時のみの利用や、使用量の制限のため、各利用者の日ごとの利用量のばらつきはブロードバンドより大きくなります。そのため、1週間分のデータから1日平均を求めると、1日単位で見た場合より利用者間のばらつきは小さくなります。1日単位で同様の分布を描くと、分布の山が少し低くなり、その分両側の裾が持ち上がりますが、基本的な分布の形や最頻出値はほとんど変わりません。

表-1は、ブロードバンド利用者の1日のトラフィック量の平均値と中間値、分布の山の頂点にある最頻出値の推移を示します。分布の山に対して頂点が少しずれている場合は、最頻出値は分布の山の中央に来るように補正しています。今回、いずれの数値も大きな伸びとなっています。分布の最頻出値を2019年と2020年で比較すると、INでは89MBから158MBに、OUTでは1995MBから3162MBに増えており、伸び率で見ると、INで1.8倍、OUTで1.6倍となっています。一方、平均

	IN (MB/day)			OUT (MB/day)		
年	平均値	中間値	最頻出値	平均値	中間値	最頻出値
2007	436	5	5	718	59	56
2008	490	6	6	807	75	79
2009	561	6	6	973	91	100
2010	442	7	7	878	111	126
2011	398	9	9	931	144	200
2012	364	11	13	945	176	251
2013	320	13	16	928	208	355
2014	348	21	28	1124	311	501
2015	351	32	45	1399	443	708
2016	361	48	63	1808	726	1000
2017	391	63	79	2285	900	1259
2018	428	66	79	2664	1083	1585
2019	479	75	89	2986	1187	1995
2020	609	122	158	3810	1638	3162

表-1 ブロードバンド個人利用者の1日のトラフィック量の平均値と最頻出値の推移

	IN (MB/day)			OUT (MB/day)		
年	平均値	中間値	最頻出値	平均値	中間値	最頻出値
2015	6.2	3.2	4.5	49.2	23.5	44.7
2016	7.6	4.1	7.1	66.5	32.7	63.1
2017	9.3	4.9	7.9	79.9	41.2	79.4
2018	10.5	5.4	8.9	83.8	44.3	79.4
2019	11.2	5.9	8.9	84.9	46.4	79.4
2020	10.4	4.5	7.1	79.4	35.1	63.1

表-2 モバイル個人利用者の1日のトラフィック量の平均値と最頻出値

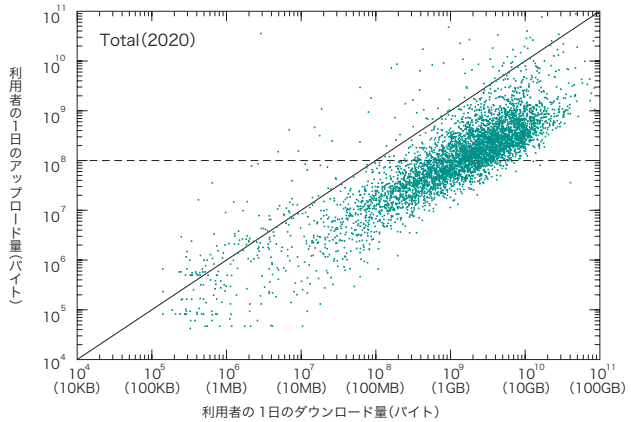


図-4 ブロードバンド利用者ごとのIN/OUT使用量

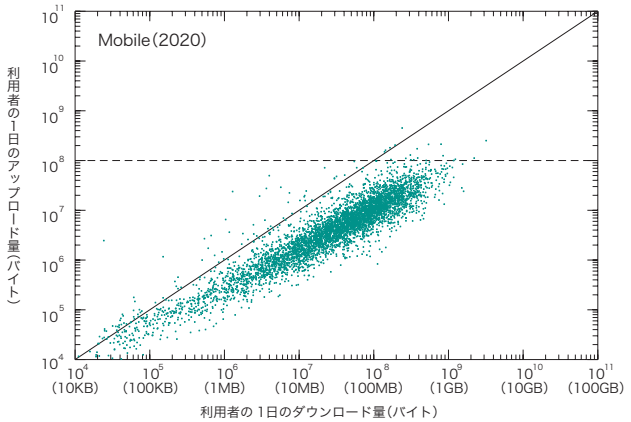


図-5 モバイル利用者ごとのIN/OUT使用量

値はグラフ右側のヘビーユーザの使用量に左右されるため、2020年には、INの平均は609MB、OUTの平均は3810MBと、最頻出値よりかなり大きな値になります。2019年には、それぞれ479MBと2986MBでした。

モバイルでは、ヘビーユーザが少ないため、平均と最頻出値が近い値になります。表-2に示すように、今回はすべての項目で減少しています。2020年の最頻出値は、INで7MB、OUTで63MBで、平均値は、INで10MB、OUTで79MBです。2019年の最頻出値は、INで9MB、OUTで79MB、平均値は、INで11MB、OUTで85MBでした。

図-4及び図-5では、利用者5,000人をランダムに抽出し、利用者ごとのIN/OUT使用量をプロットしています。X軸はOUT（ダウンロード量）、Y軸はIN（アップロード量）で、共にログスケールです。利用者のIN/OUTが同量であれば対角線上にプロットされます。

対角線の下側に対角線に沿って広がるクラスタは、ダウンロード量がひと桁多い一般的なユーザです。ブロードバンドでは、以前は右上の対角線上あたりを中心に薄く広がるヘビーユーザのクラスタがはっきり分かりましたが、今では識別ができなくなっています。また、各利用者の使用量やIN/OUT比率にも大きなばらつきがあり、多様な利用形態が存在することが窺えます。モバイルでも、OUTがひと桁多い傾向は同じですが、ブロードバンドに比べて利用量は少なく、IN/OUTのばらつきも小さくなっています。ブロードバンド、モバイル共に2019年との違いはほとんど分かりません。

図-6及び図-7は、利用者の1日のトラフィック量を相補累積度分布にしたものです。これは、使用量がX軸の値より多い利用者の、全体に対する割合をY軸に、ログ・ログスケールで示したもので、ヘビーユーザの分布を見るのに有効です。グラフの右側が直線的に下がっていて、べき分布に近いロングテールな分布であることが分かります。ヘビーユーザは統計的に分布しており、決して一部の特殊な利用者ではないと言えます。

モバイルでも、OUT側ではヘビーユーザはべき分布していますが、IN側では直線的な傾きが崩れていて、大量にアップロードするユーザの割合が大きくなっています。今年は、INの分布の右端が更に右側に伸びていて、一部の大量アップロードするユーザのアップロード量が一層増えています。

利用者間のトラフィック使用量の偏りを見ると、使用量には大きな偏りがあり、結果として全体は一部利用者のトラフィックで占められています。例えば、ブロードバンド上位10%の利用者がOUTの50%、INの76%を占めています。更に、上位1%の利用者がOUTの16%、INの50%を占めています。昨年と比べると、偏りは少し減少しています。モバイルでは、上位10%の利用者がOUTの48%、INの53%を、上位1%の利用者がOUTの13%、INの23%を占めています。こちらは昨年のレポートより偏りが少し大きくなっています。

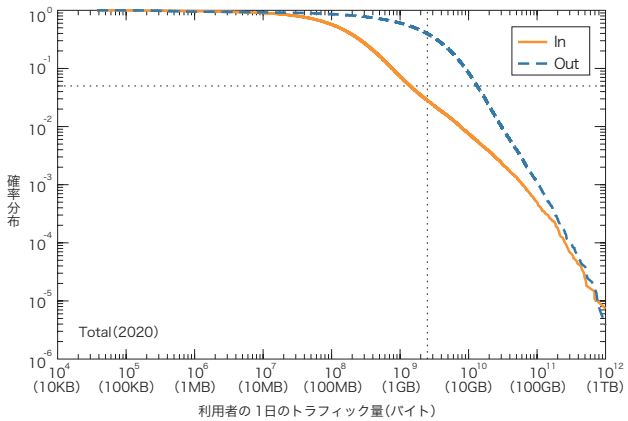


図-6 ブロードバンド利用者の1日のトラフィック量の相補累積度分布

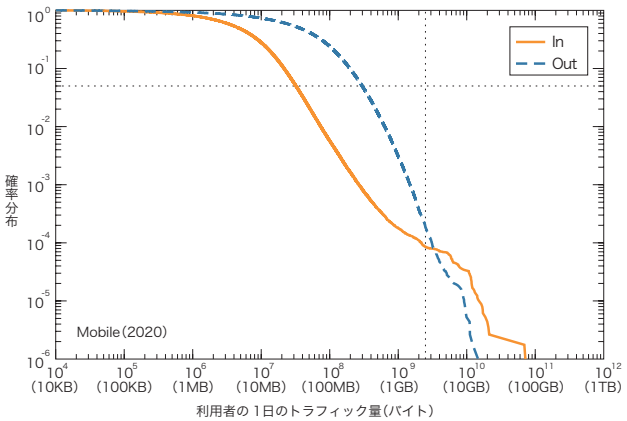


図-7 モバイル利用者の1日のトラフィック量の相補累積度分布

1.4 ポート別使用量

次に、トラフィックの内訳をポート別の使用量から見ていきます。最近では、ポート番号からアプリケーションを特定することは困難です。P2P系アプリケーションには、双方が動的ポートを使うものが多く、また、多くのクライアント・サーバ型アプリケーションが、ファイアウォールを回避するため、HTTPが使う80番ポートを利用します。大まかに分けると、双方が1024番以上の動的ポートを使っていればP2P系のアプリケーションの可能性が高く、片方が1024番未満のいわゆるウェルノウンポートを使っていれば、クライアント・サーバ型のアプリケーションの可能性が高いと言えます。そこで、TCPとUDPで、ソースとデスティネーションのポート番号の小さい方を取り、ポート番号別の使用量を見てみます。

表-3はブロードバンド利用者のポート使用割合の過去5年間の推移を示します。2020年の全体トラフィックの77%はTCPです。HTTPSのTCP443番ポートの割合は、52%で前回とほぼ同じです。HTTPのTCP80番ポートの割合は20%から17%に減っています。QUICプロトコルで使われるUDP443番ポートは、8%から11%に増えていて、HTTPの減った分とQUICの増えた分がほぼ同じです。

減少傾向のTCPの動的ポートは、2020年には7%にまで減りました。動的ポートでの個別のポート番号の割合は僅かで、最大の8080番でも0.4%となっています。また、Flash Playerが利用する1935番が0.4%ありますが、これら以外のトラフィックは、ほとんどがVPN関連です。

表-4はモバイル利用者のポート使用割合です。全体的にはブロードバンドの数字に近い値となっています。これは、スマートフォンでもPCと同様のアプリケーションを使うようになってきたことに加え、ブロードバンドにおけるスマートフォンの利用割合が増えているからだと思います。

図-8は、ブロードバンド全体トラフィックにおける主要ポート利用の週間推移を、2019年と2020年で比較したものです。TCPポートの80番、443番、1024番以上の動的ポート、UDPポート443番の4つに分けてそれぞれの推移を示しています。グラフでは、ピーク時の総トラフィック量を1として正規化して表しています。2019年と比較すると、コロナ禍での在宅時間の増加に伴い、平日昼間のトラフィックが大きく増えているのが分かります。全体のピークは19時から23時頃です。

year	2016	2017	2018	2019	2020
protocol port	(%)	(%)	(%)	(%)	(%)
TCP	**82.8**	**83.9**	**78.5**	**81.2**	**77.2**
(< 1024)	69.1	72.9	68.5	73.3	70.5
443(https)	30.5	43.3	40.7	51.9	52.4
80(http)	37.1	28.4	26.5	20.4	17.2
993(imaps)	0.1	0.2	0.2	0.3	0.2
22(ssh)	0.2	0.1	0.1	0.2	0.2
182	0.3	0.3	0.3	0.2	0.2
(>= 1024)	13.7	11.0	10.0	7.9	6.7
8080	0.2	0.3	0.3	0.5	0.4
1935(rtmp)	1.5	1.1	0.7	0.3	0.4
UDP	**11.4**	**10.5**	**16.4**	**14.1**	**19.4**
443(https)	2.4	3.8	10.0	7.8	10.5
8801	0.0	0.0	0.0	0.0	1.1
4500(nat-t)	0.2	0.2	0.2	0.3	0.6
ESP	**5.8**	**5.1**	**4.8**	**4.4**	**3.2**
GRE	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**
IP-ENCAP	**0.2**	**0.3**	**0.2**	**0.2**	**0.1**
ICMP	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**

表-3 ブロードバンド利用者のポート別使用量

year	2016	2017	2018	2019	2020
protocol port	(%)	(%)	(%)	(%)	(%)
TCP	**94.4**	**84.4**	**76.6**	**76.9**	**75.5**
443(https)	43.7	53.0	52.8	55.6	50.7
80(http)	46.8	27.0	16.7	10.3	7.4
993(imaps)	0.5	0.4	0.3	0.3	0.2
1935(rtmp)	0.3	0.2	0.1	0.1	0.1
UDP	**5.0**	**11.4**	**19.4**	**17.3**	**18.0**
443(https)	1.5	7.5	10.6	8.3	9.3
4500(nat-t)	0.2	0.2	4.5	3.0	1.8
8801	0.0	0.0	0.0	0.0	1.4
1701(12tp)	1.0	0.0	0.0	0.4	0.9
12222	0.1	0.1	2.3	3.4	0.8
ESP	**0.4**	**0.4**	**3.9**	**5.8**	**6.4**
GRE	**0.1**	**0.1**	**0.1**	**0.0**	**0.1**
ICMP	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**

表-4 モバイル利用者のポート別使用量

図-9のモバイルでは、トラフィックの大半を占めるTCP80番ポートと443番ポート、UDP443番ポートについて推移を示します。2019年と比較すると、モバイルはほとんど変化がありません。ブロードバンドに比べると、朝から夜中までトラフィックの高い状態が続きます。平日には、朝の通勤時間、昼休み、夕方17時頃から22時頃にかけての3つのピークがあり、ブロードバンドとは利用時間の違いがあることが分かります。

1.5 まとめ

ここ数年、トラフィック量は緩やかな伸びが続いていましたが、今回は、新型コロナウイルス感染拡大によってインターネットの利用に大きな変化が起こりました。リモートワークが急拡大し、学校の授業もオンラインで行われるようになり、平日昼間のトラフィック量が大きく増えました。また、オンライン会議ツールが急速に普及して、飲み会などの交流や子供達の習い事や部活にも使われるようになってきています。

トラフィック量的には、コロナ禍の影響が一番大きく出ていたのは5月で、今回報告した6月には少し落ち着いてきていましたが、それでも昨年の同時期よりブロードバンドのダウンロードが34%も増えています。利用者ごとの利用量を見ても、ブロードバンドは在宅時間が増えて大きく伸び、モバイルは外出自粛で減りました。

今年は、本来ならばオリンピック・パラリンピックの開催に伴ったネットサービスの展開による利用動向の変化を予想していたのですが、新型コロナウイルス感染拡大の影響で予想もしなかった形でインターネット利用状況に変化が起こりました。6月以降トラフィック量的には少し落ち着いてきてはいますが、リモートワークやビデオ会議の利用などは定着してきていて、元のレベルに戻ることはないでしょう。現時点ではまだまだ先が見えない状況が続いているので、引き続き状況を注視していきたいと思います。

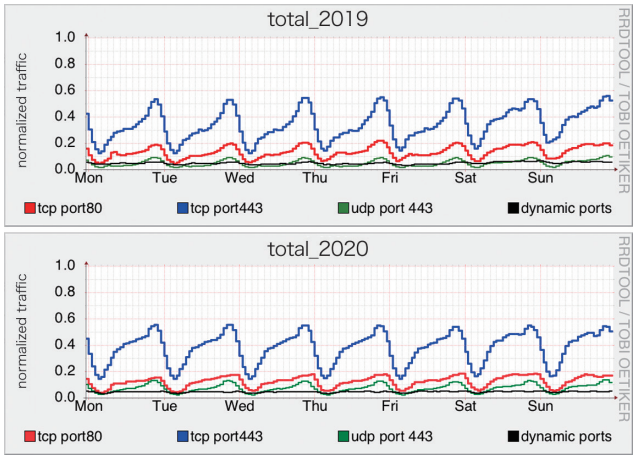


図-8 ブロードバンド利用者のポート利用の週間推移
2019年(上)と2020年(下)

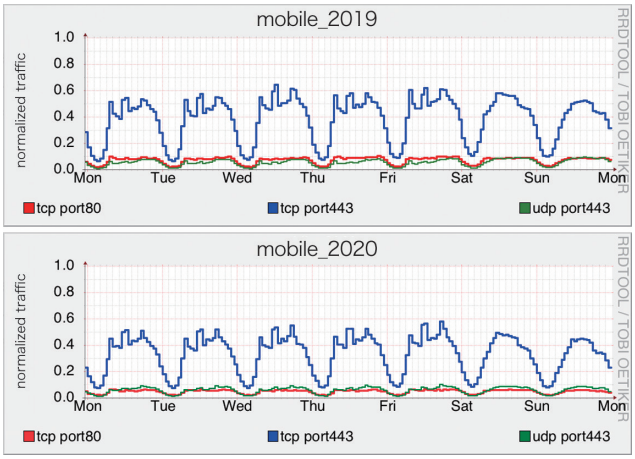


図-9 モバイル利用者のポート利用の週間推移
2019年(上)と2020年(下)

執筆者:
長 健二朗（ちょう けんじろう）
株式会社IIJ イノベーションインスティテュート 技術研究所所長。

5G時代のMVNOの在り方 ～VMNO構想の実現に向けた取り組み

2.1 5Gに向けた助走

IIJは2008年にセルラー通信網(当初はW-CDMA、その後2012年からはLTE)を利用したMVNO事業を開始して以来、一貫してこの分野のフロントランナーを務めてきました。この間、MVNOを巡る市場環境は大きく様変わりしましたが、IIJは法人向け、個人向け、MVNE事業、IoT/M2M、そしてフルMVNOと、多様かつ先進的なMVNOビジネスを展開し、多くの皆様にご利用いただいています。現在、それら回線数の合計は300万回線を超えてなお成長を続けており、名実共に日本最大のMVNOとなりました。

そんな中、MVNOの競争環境は日に日に激しさを増しています。特にスマートフォンの販売方法に関する直接的規制が年々厳しくなる中、これまでのような多額のキャッシュバックや2年の期間拘束契約を前提とした高価格帯料金プラン・ハイエンドモデル端末を提供するMNOと、いわゆる「ノーフリル」型料金プラン・ミドルクラスやローエンド端末中心のMVNOによる、垂直型の市場構造が崩壊し、競争は多角的かつ多面的な様相を見せています。MNOのサブブランドや、第4のMNOである楽天モバイルの台頭が今後も進んでいく中、MVNOの中には既に収益の確保に苦しむケースが散見されます。中には市場からの撤退を選ばざるを得ないMVNOも見られる状況です。このような流れがなぜ生じているのでしょうか。

MVNOは、MNOが提供する限られた通信サービスのみ提供することが可能なビジネスです。MNOが、技術的に実現可能な通信サービス全ての中から投資対効果(収益性)や他社との差別化を考慮し、どのような通信サービスを提供するかを主体的に選択できるのとは対照的に、MVNOはMNOのネットワークを活用する限りにおいて、その制約の元で選択肢を考慮する以外にありません。1990年代の2Gから、3G、そして4G LTEとセルラー通信技術の世代が進んでいく中、当初は従量制の音声通話のみだった携帯電話の通信サービスは、パケット通信、VoLTE、音声定額プランやパケット定額プラン、キャリアアグリゲーション、LPWAと様々な進化を遂げてきましたが、これらのMVNOに対する提供はあくまでMNOの提示する技術的・経済的な条件にのみ基づくものであり、MVNOは本質的に差別化を図っていくことが難しいといえるのです。

とはいえ、MVNOにもできることはあります。それは、MNOの通信設備の一部を外に切り出し、自らその設備を運用することで、その設備を用いて実現可能な通信サービスを主体的に提供していくという道です。このような「切り出し」を「アンバンドル」と呼びます。MVNOの歴史はアンバンドルを広げていく歴史でもありました。パケット交換機[*1]のアンバンドル、いわゆる「レイヤー2接続」は、2008年の総務大臣裁定で認められ、その後MNO3社に対し義務化されていきました。SIMカードを管理す

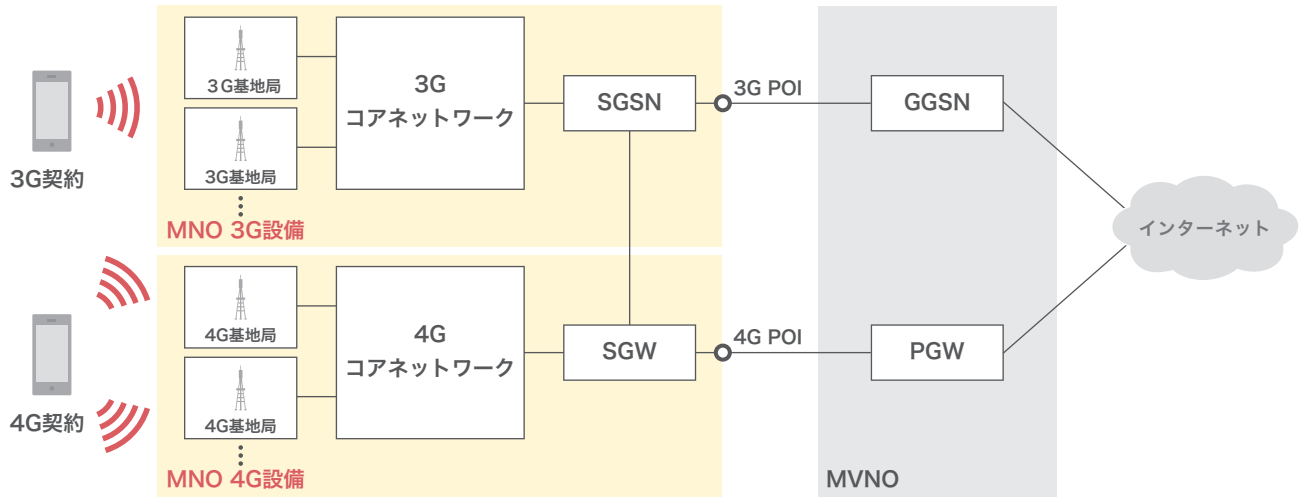


図-1 パケット交換機のアンバンドル(レイヤー 2接続)のイメージ

＊1　3Gでは「GGSN」(Gateway GPRS Support Node)、4G LTEでは「PGW」(Packet GateWay)と呼ばれる設備。

る加入者管理装置*2のアンバンドル、いわゆる「フルMVNO」は、MNO3社に対し義務化まではされていないものの、開放することが望ましい機能であるとガイドラインで指定されたため、2018年のIIJを皮切りに、HLR/HSSのアンバンドルを受け自ら運用し、他のMVNOでは実現できない様々なビジネスを提供するフルMVNOが登場しています。フルMVNOとしての新サービス開発への弊社の取り組みについては、IIR Vol.38*3にてご紹介していますので、ぜひこちらもご覧ください。

しかし、これから本格化してくる5G時代を見据えたとき、我々はこれまでのビジネスの延長ではなく、新たな岐路に立っているといえるでしょう。5G時代の当初に登場する、4Gの設備に依存する形のNSA*4と呼ばれる5Gの実装では、これまでの4Gから設備面は大きく変わりませんが、その後に登場する、4Gの設備に依存しないSA*5と呼ばれる実装が提供される頃には、MNOの5Gネットワークは高度に仮想化していくことが期待されています。5Gの目指す多様なエンドツーエンドのQoS*6、すなわち超高速モバイルブロードバンド、同時多数接続、超低遅延・高信頼性通信の3つを効率良く実現するためには、仮想化技術の導入と、それによるネットワークの水平的な分割、すなわちネットワークスライシングの実現が必要不可欠だからです。

しかし、MVNOから見れば大きな疑問が未解決のものとして残ります。5G SA時代の仮想化ネットワークにおいても、これまでのようにアンバンドルによる設備の切り出しは機能し続けるでしょうか。機能しないとすれば、どのようにすれば我々MVNOは差別化を図っていくことができるのでしょうか。

2.2 5GとMVNO

5G SAにおけるアンバンドルの可能性を考えたとき、大きな課題が2つあることに気が付きます。その1つはネットワークの分割そのものに起因する問題です。アンバンドルは、事業者間に責任分界点(POI*7)を置きネットワークを垂直的に分割する手法ですが、これがネットワークスライシングと相性が悪いのです。すなわち、5Gで求められる様々なエンドツーエンドのQoS確保のためにネットワークスライシング、すなわち水平的な仮想ネットワークの分割を導入するにもかかわらず、MVNOがその一部の機能のみを更に物理的に切り離せば、本来達成しなければならないQoSの実現に支障が生じる可能性が想定されます。

もう1つの課題は運用面です。POIにおける事業者間の技術的インタフェース仕様は一般に標準化されていることが期待されます。しかし、それだけでなく、POIを挟んだ両側の運用主体が異なることから、その運用における制約も非常に大きいのです。仮にPOIにおける技術的インタフェース仕様を満たしているとしても、何かの設定を変えたり、新たに機能を付け加えたりすることは、POIを挟む事業者双方の合意なく行うことはできません。これまでの3Gや4G LTEの世界では、レイヤー2接続にしろ、フルMVNOにしろ、POIを構築した後にその設定を変えるということはそんなに頻繁に起きることではありませんでした。ですので、このような制約があってもそれ程不自由なくビジネスを進められてきました。しかし、5Gでは利用者の要求に応える通信サービスを提供するために様々なQoSを実現するスライス(仮想化されたコアネットワーク)をダイナ

*2 3Gでは「HLR」(Home Location Register)、4Gでは「HSS」(Home Subscriber Server)と呼ばれる設備。
*3 Internet Infrastructure Review(IIR)Vol.38 フォーカス・リサーチ(1)「フルMVNOとは何か、IIJはなぜフルMVNOを目指すのか」(https://www.iij.ad.jp/dev/report/iir/038/02.html)。
*4 Non-StandAloneの略。
*5 StandAloneの略。
*6 Quality of Serviceの略。
*7 Point Of Interfaceの略。

ミックに運用していく必要があります。このような5Gの柔軟性を、これまでの手法、すなわちPOIを挟んで一部の機能を切り出すアンバンドルで実現することは、運用面においても非常に難しいと言わざるを得ません。

2.3 VMNO構想とは

この2つの5G SA時代の課題を解決することを目的に、IIJ及びMVNOの業界団体である(一社)テレコムサービス協会 MVNO委員会は、5G時代における新たな仮想通信事業者のコンセプトとして、VMNO構想を提唱しています。そのオリジナルの考え方は、欧州の1つの報告書に端を発しています。

欧州のシンクタンクであるCERREが2017年3月に発行した報告書[*8]では、欧州における5Gのリーダーシップを確保するための道筋として、これまでの4Gまでのやり方を5Gでも続けていく「エボリューション(進化)」と、これまでのアプローチを大きく変えるべきとする「レボリューション(革命)」の2つのシナリオが提示されました。このうち、後者の中心となるのが「VMNO」、すなわちVirtual MNO (仮想MNO)です。多種多様な産業に対しカスタマイズされた通信ソリューションを提供するという5Gのミッションを遂行するためには、MNOの数は少なすぎ、また物理的なインタフェースに縛られたMVNOはそこまでのビジネスの自由度を持ち得ないことを指摘した上で、5Gのネットワークスライシングを管理するAPIをMNOが外部に開放することで誕生する多数の「VMNO」が、MNOと同等の自由度により産業に特化した5Gソリューションを展開していく、とするシナリオです。

その後、2019年9月に同じCERREが発行した白書[*9]では、更に踏み込んで、4G時代までのフルMVNOは5Gの仮想化ネットワークの中では不可能になる可能性があるとし、VMNOコンセプトの実現を呼びかけました。図-2は、VMNO (この図及びこのレポートでは便宜的に「ライトVMNO」と呼ぶ)とホストMNOの間の、想定される構造を示します。

現行世代のアンバンドルが、コアネットワークをPOIでMNO側、MVNO側に分割しているのに対し、ライトVMNOのイメージではコアネットワークそのものはMNOにより統合的に運用される点が大きく異なります。ライトVMNOが有するのは運用やビジネスを司る業務システムOSS/BSS[*10]のみで、これがMNOのネットワークにあるAPIを用いてスライスにアクセスします。

このような構造を取ることにより、MNOの提供するAPIを通じ、ライトVMNOはMNOの仮想化基盤に置かれた仮想化コアネットワーク、すなわちスライスを管理することが可能となります。APIの組は2つ、1つは個々のスライスが実現しているコアネットワーク、すなわち利用者に提供される通信サービスのQoSを管理するためのもので、もう1つは、スライスの追加や、不要となったスライスの削除など、それらスライス自体を管理するためのものとなります。

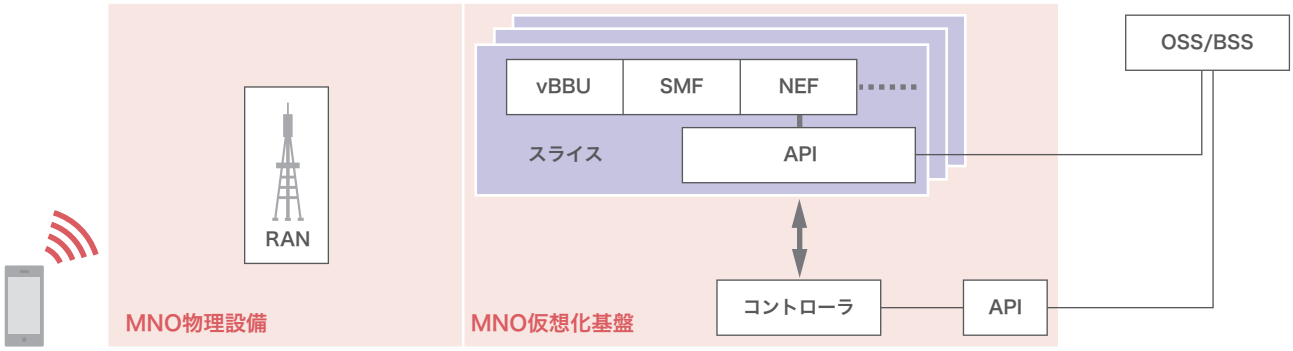


図-2 ライトVMNOの想定される構造

*8 "Towards the successful deployment of 5G in Europe"(https://www.cerre.eu/sites/cerre/files/170330_CERRE_5GReport_Final.pdf)。
*9 Ambitions For Europe2024(https://cerre.eu/sites/cerre/files/cerre_whitepaper_ambitionsforeurope2024.pdf)。
*10 Operation Support System/Business Support Systemの略。

IIJ及びテレコムサービス協会の提唱するもう1つのVMNOのモデルが「フルVMNO」です。ライトVMNOが、MNOの提供する仮想化基盤の上でビジネスを展開するのに対し、フルVMNOは自身で仮想化基盤を保有するところが大きく異なります。図-3にフルVMNOの想定される構造を示します。

ライトVMNOとフルVMNOの違いは仮想化基盤のオーナーシップです。ライトVMNOがOSS/BSSを除きホストMNOの設備に依存するのに対し、フルVMNOは無線部分を除きMNOの設備から独立しています。この違いにより、フルVMNOは、ホストMNOからの更なる技術的、運用的な独立性を保ち、他の無線通信事業者とのコラボレーションの可能性を有します。この種の独立性は、現行世代のフルMVNOが有するものです。フルVMNOは、現行世代のフルMVNOが未だ成し得ていない「複数の無線網への乗り入れ」の実現にチャレンジすることになるでしょう。

これらVMNO構想に関し、テレコムサービス協会MVNO委員会から提案を受けた総務省の研究会「モバイル市場の競争環境に関する研究会」は、2020年2月にまとめられた報告書において、これら2つのVMNOモデルの双方について、来るべき5G SA時代における仮想通信事業者のコンセプトとして検討を進めることが適切であるとしました。これにより、VMNO構想は今後の仮想通信事業者の在り方として最も有力な選択肢となったのです。

2.4 VMNOの実現によるメリット

このように着実に進みつつあるVMNO構想ですが、どのようなメリットをもたらすのでしょうか。

CERREは白書の中で、VMNOによってもたらされる新たな市場構造は、B2B市場とB2C市場の双方に等しく、リテールレベルにおける活発な競争をもたらす可能性を秘めていると提唱しています。それは、日本を含む多くの国で、限られた無線周波数資源から自ずと数が限られ、よりマクロに見れば合併や統合により緩やかに数を減らしつつあるMNOと比較して、多数のVMNOの登場が期待できるためでしょう。VMNOは現行世代のMVNO同様、周波数の割り当てを受けない仮想移動通信事業者であるため、限られた周波数資源といった自然的条件による市場参入数の制約はありません。しかも、現行世代のMVNOのようにMNOからの機能の提供条件やアンバンドルの可否に事業の自由度が縛られることはなく、より広い選択肢から自らの顧客にとって必要な機能を選び、求められるQoSの通信サービスを提供することが可能です。このようなVMNOが市場に存在すれば、競争は自ずと活発になり、5G SA時代において利用者が求めるサービスをより一層得やすくなることが考えられます。

日本においても、テレコムサービス協会MVNO委員会は、高い自由度を持つVMNOの存在が、革新的なソリューションの創出を加速させることを指摘しています。この恩恵は、例えば中

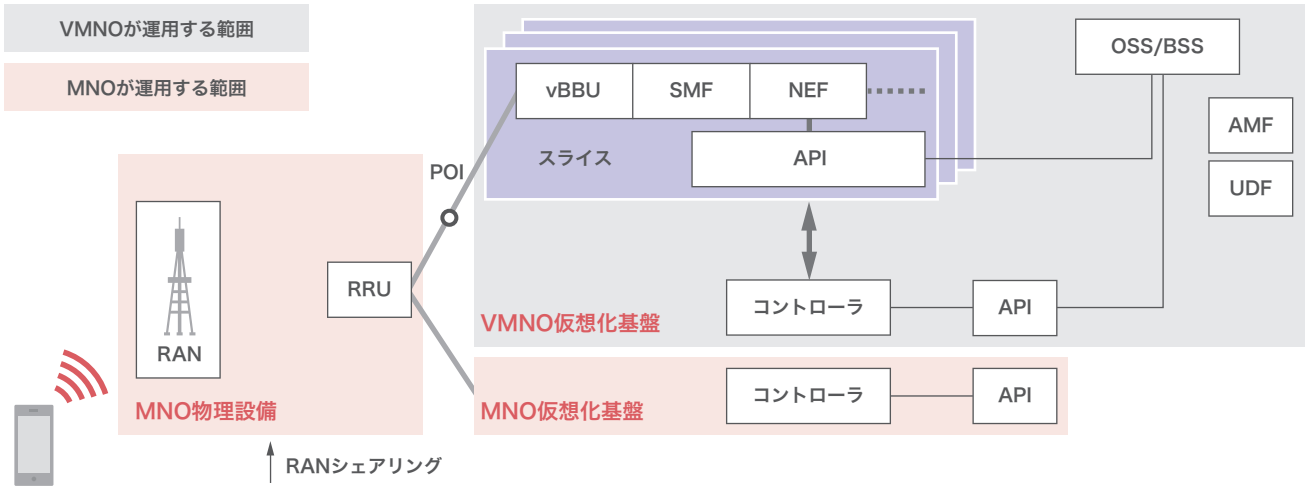


図-3 フルVMNOの想定される構造

小企業や地方など5Gの普及が比較的遅れることが見込まれる市場や地域における5G導入の問題を解決することになるでしょう。

更に、特定のMNOのインフラストラクチャーに依存しないコアネットワークを有するフルVMNOについては、総務省が強力に推し進めている「ローカル5G」[*11]の普及においても重要な役割を果たすことが期待されています。フルVMNOは、SIM、端末、仮想化基盤とコアネットワーク、OSS/BSSなど、ローカル5G事業者が必要とする全てのコンポーネントを有します。しかも、これらが特定の無線網の運用から独立していることから、しがらみなく様々な無線網を利活用できるというビジネス上のポジションを持ち、ローカル5G事業者の要求事項を、特定の無線網にこだわらずに満たすことができる点で他の追随を許さない立場にいます。IIJでは、フルVMNOがいわば「ローカル5Gイネイブラー」となることで、スタジアム、病院、ホテルや工場のオーナーなどに代表される、自らのサイトの中での高品質かつ安価なプライベートのセルラーコネクティビティの実現を望んでいるローカル5G事業者との間で、全く新しいタイプの通信事業を実現することができると考えています。

2.5 VMNOの実現に向けた課題

一方で、VMNOの実現には課題も多くあります。VMNOを実現するためには、技術的、ビジネス的、そして制度的な対応が必要となるでしょう。それぞれを詳しく見ていきます。

技術面では、ライトVMNO、フルVMNOのそれぞれで異なるハードルが存在します。ライトVMNOではAPIの標準化が課題となります。ライトVMNOが必要とするAPIの技術インタフェース条件が標準化されることで、ライトVMNOの実現は容易となります。このような標準化がなされなかったり、あるいは不足していたりする場合には、ライトVMNOは都度必要なAPIや機能の開発をMNOに求めることになり、ライトVMNOの実現は非常に難しいものとなるでしょう。一方、フルVMNOについてはRANシェアリングの円滑な実現が挙げられるでしょう。複数のコアネットワークで1つの無線ネットワークをシェアリングするRANシェアリングは、既に国内の一部のMNOは運用をしており、今後の5G展開に向けてコスト面での切り札としても位置付けられています。現在のところはまだ1つのMNOグループ内に閉じたRANシェアリングとなっているのが現状ですが、今後、5Gの展開に向けMNOグループの垣根を超えたRANシェアリングの実現に進むとなれば、その枠組みに参加することになるであろうフルVMNOにとっても好機となり得ます。このような標準化の取り組みは必ずしも日本国内で行われるものではないため、グローバルでの課題認識の共通化も必要となるでしょう。国際連合の特別機関の一部として電気通信関連の標準化を進めるITU-Tの第3研究委員会（料金・会計原則と国際的な通信/ICTの経済・政策）のアソシエイトメンバーであるIIJは、既にVMNO構想を含む寄書を同委員会に提出しており、今後、国際的な認知、理解の向上に向け議論が進んでいく見込みです。

*11 地域や産業の個別のニーズに応じて地域の企業や自治体などの様々な主体が自らの建物内や敷地内でスポット的に柔軟に構築できる5Gシステム。28GHz帯（ミリ波）の100MHz帯域は既に導入済みで、今後28GHz帯の残る800MHz帯域、及びより周波数が低くエリアを構築しやすい4.6GHz帯の300MHz帯域について2020年中の制度化を目指している。

ビジネス面では、MNOとVMNO双方の利害のすり合わせが必要となります。VMNOは、MNOの構築する5G設備(基地局・コアネットワーク)の収益性を高め、新しいソリューション開発によって5Gの普及に貢献するという点においてはMNOのパートナーといえますが、反面、ソリューション営業の現場では商売敵ともなり得ます。このような相克はこれまでもMVNOがMNOとの間で長く直面し続けた問題ですが、5Gに向けても引き続き良いパートナーシップの確立に向けた水面下、水面上での動きを継続していく必要があります。

制度面では、これまでの、事業者間接続という1985年の通信自由化以来の他者設備利用モデルが大きく転換することが最大の課題です。現在の電気通信事業法では、大きく「(事業者間)接続」「卸役務」の2つの他者設備利用モデルが存在していますが、特にMVNOのデータ通信のコンテキストでは、MNO側の義務が大きい「接続」がベースとしてあり、総務省令に基づいて計算されるデータ通信のネットワークの賃借料、すなわち接続料を「卸役務」でもそのまま準用することで、接続であっても卸役務であっても同じ条件でMNOの設備利用が可能となる構図でした。しかし、特にライトVMNOにおいてはPOIすなわち事業者間接続における電気的接続点はなく、また電気的接続点があることが想定されるフルVMNOにおいても電気通信事業法上の位置付け、接続料をどう考えるべきかは未だ議論がされていません。接続料を、卸役務のみを念頭に民間協議に任せるのか、制度的にその計算方法や上限などを決めるのかなど、これからの議論に委ねられています。

2.6 おわりに

VMNOについては、前提となる5G SA時代のネットワーク仮想化自体がまだ将来の話であり、すぐに実現するビジネスモデルではありません。しかし、IIJにとってのフルMVNOがそうであったように、全く新しいビジネスモデルを構築するためには数年単位での時間が必要であることから、早期に議論を開始することは必要不可欠であると考えています。IIJでも、業界団体経由での取り組みのみならず、我々にできる取り組みは進めています。全く新しい事業形態を実現するのですから決して容易なことではないとはいえ、IIJはVMNO構想実現に向け引き続き活動を進めていきます。

執筆者:
佐々木 太志(ささき ふとし)
IIJ MVNO事業部 ビジネス開発部 担当部長。
2000年IIJ入社、以来ネットワークサービスの運用・開発・企画に従事。
特に2007年にIIJのMVNO事業の立ち上げに参加し、以後一貫して法人向け、個人向けMVNOサービスを担当。
MVNOの業界団体である一般社団法人テレコムサービス協会MVNO委員会にもメンバーとして参加。

Splunkによる日本語文章解析処理

3.1 はじめに

数百万アカウントを収容する大規模メールサービスとなるIIJ xSPプラットフォームサービス/Mailでは、大量蓄積するログからの有用な情報抽出・システム解析・迷惑メール送信者と戦うためにSplunk[*1]を導入しました。

導入当初はログ検索を中心に利用していましたが、昨今はSplunk Machine Learning Toolkit（図-1）[*2]を用いたスパム検知自動化、サービス運用の効率化など、幅広くSplunkを活用しています。

今回はSplunkの導入経緯から始まり、Splunk Deep Learning ToolkitのNLPに日本語処理機能を追加拡張しSplunk社にフィードバック・マージされたお話と、これを用いたテキストマイングについて紹介します。

3.2 Splunk導入経緯

IIJ xSPプラットフォームサービス/Mailでは、顧客サポートセンター向けの機能としてメール配送検索、個々のメールの配送経路表示、WEBメール、POP/IMAP/SMTP認証ログの検索機能など、サポート窓口のスタッフがエンドユーザからの問い合わせに対してログを調査する機能を、ElasticSearchを用いて実装しています。またこの他にIIJ社内のサービス運用ツールとして、サービス立ち上げ当初は大量打ち込みを行っているユーザの特定、エラー検出、顧客向けレポート作成などにElasticSearch、Kibanaを活用していました。

IIJ xSPプラットフォームサービス/Mailでは、更なるサービス品質向上を目的に、スパム検知精度を上げるためMachine Languageアルゴリズムの導入検討を進めた際、ElasticSearch、Kibanaに限界を感じていることもあり、

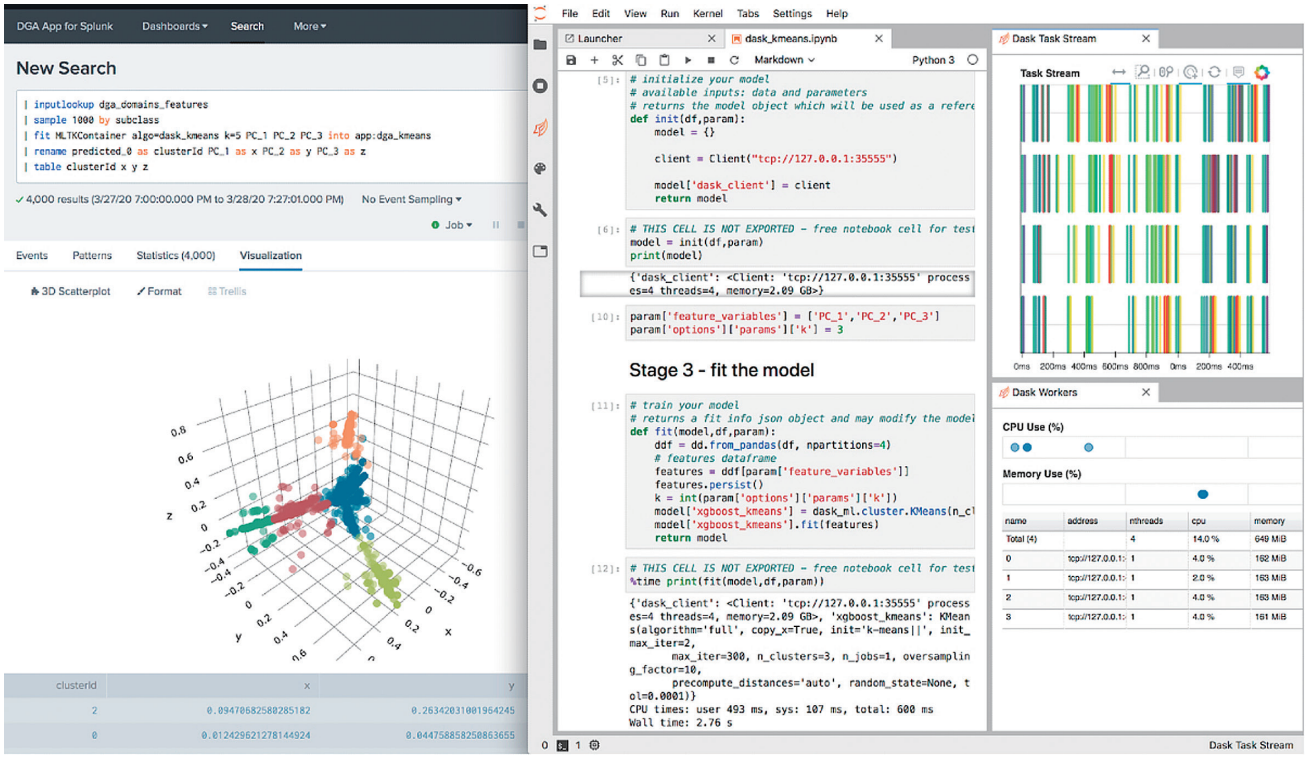


図-1 Splunk Machine Language Toolkitイメージ

＊1　Splunk Enterprise：統合ログ解析・管理ツールビッグデータ分析ソフトウェア（https://www.splunk.com/ja_jp/software/splunk-enterprise.html）。
＊2　Splunk Machine Language Toolkit（https://www.splunk.com/en_us/blog/machine-learning/deep-learning-toolkit-3-1-examples-for-prophet-graphs-gpus-and-dask.html）。

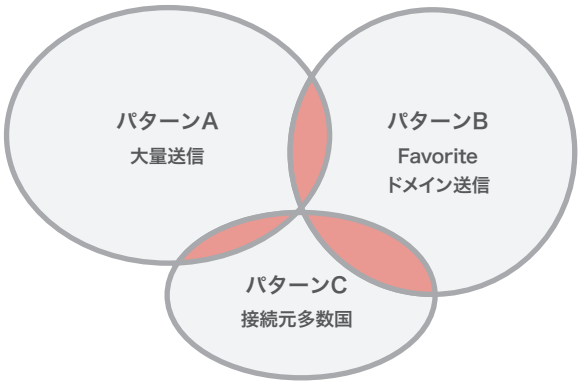
Splunkを導入するに至りました。Splinkは様々な目的に最適化されたプラグイン、可視化Appが豊富(無償/有償)でスピード感のある開発が期待できる上、ElasticSearchと比較して圧倒的なシステム安定性と保守のしやすさがあり、無償のMachine Learning Toolkit/Deep Learning Toolkitが魅力的であることがその理由です。

3.3 Splunkを活用したスパム検知

Machine Learningを用いて精度を上げるためにはアルゴリズムの選択の他、解析軸の選択、アルゴリズムのパラメータ調整、学習、モデルの検証の繰り返し実行が必要ですが、Splunk Machine Learning Toolkit/Deep Learning Toolkitでは、これらがシームレスに実行できるUI環境が提供されており、短期間でアルゴリズムを評価し、モデル精度を上げることができました。

スパムは様々な手法を使って正規ユーザの中に紛れるように活動しています。またはスパムによりActivityの特徴が異なるため、総合的に見て判別する必要があります(図-2)。

IIJ xSPプラットフォームサービス/Mailでは、送信元IP数、送信元国数、一定時間内における送信数、宛先ユニーク数、スパマが好んでターゲットとするドメインを主に対象として送信しているのか、それ以外のドメインに一様に送信しているのか、送信結果のエラー発生率など、複数変数の組み合わせとアルゴリズム評価を行った結果、SVM[*3]で良い結果を得ることができました。SVMは優れた認識性能を発揮する教師であり学習モデルで、n次元の超平面を扱うことができます。マージン最大化という方法で各クラスから最も遠い境界線を引くという特徴もあります。



例:左の色付き部分のように複数パターンの合致部分が真正のスパマ

図-2 スパマのActivityイメージ

3.4 日本語分析ニーズとNLP(Natural Language Processing)

サービスの様々なログを分析することによりサービス運用・運営に有用なデータを得て付加価値創造を目指してきましたが、定点で取得したスパム検体の特徴分析以外にも、ABUSE対応に困っている、Redmineのチケットを取り込んで分析しているなどの声が他部署からあり、社内においても日本語テキストデータを分析するニーズがあることが分かってきました。

ABUSEメールやRedmineチケットのテキストデータをNLPで解析することにより、人や設備などを軸とした分析を行うことで負荷や問題の集中などの早期発見が可能になります。

SplunkでMeCabを使った形態素解析が可能ですが、これだけでは大量のテキストデータの処理や高度なテキストマイニングを行うのは困難です。そこでSplunk Deep Learning ToolkitにあるNLPの利用を考えました。NLPを用いることにより、テキストデータの構文構造解析、固有表現抽出などが可能になり、大量のテキストデータを取り込みテキストマイニングが可能になるところに大きな魅力を感じました。固有表現抽出というのは、テキストから固有表現(Named Entity)を抽出

*3 SVM:Support Vector Machine。機械学習アルゴリズムの1つ。

し、更に人、組織、地名、日付や数値など、あらかじめ定義されている属性（Entity）に分類、抽出する技術です（図-3）。

検証着手当時Splunk Deep Learning ToolkitのNLPは日本語処理に対応していませんでしたので、独自拡張して日本語対応を行い、SplunkbaseというSplunkの公式ライブラリー上で公開しました。現在はSplunk Deep Learning Toolkitにマージされています。Splunk Deep Learning ToolkitのNLPを日本語対応したことにより、日本でのビジネス活用範囲が広がったことで多くの反響をいただきました。GOJAS（Splunk日本ユーザ会）のイベント講演では、100名を超える方に聴講していただくことができました（図-4）。

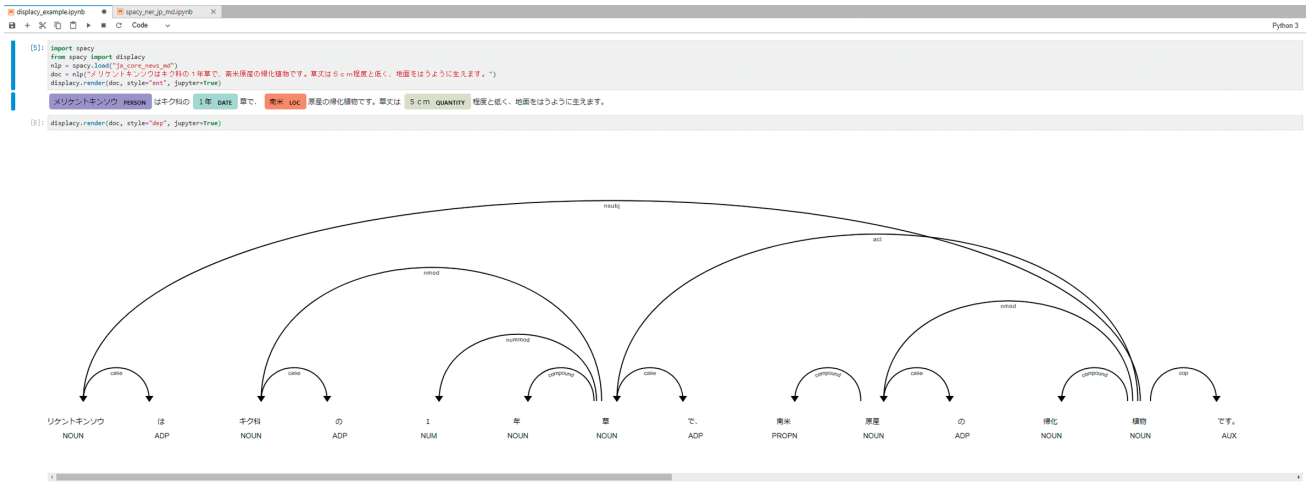


図-3 Jupyter上での固有表現抽出例

Big Thanks to the Community

Recently a DLTK user in Japan built an extension to be able to apply the Ginza NLP library on Japanese Language text and to make the NLP example work for Japanese. Luckily we were able to get his contribution merged into the DLTK 3.1 release. I'm really happy to see this community mindset and I want to thank you, Toru Suzuki-san for your contribution, ありがとうございました!

Last but not least I would like to thank so many colleagues and contributors who have helped me finish this release. A special thanks again to Anthony, Greg, Pierre and especially Robert for his continued support on DLTK and making Kubernetes a reality today!

With the upcoming .conf20 and the recently opened **'Call For Papers'** I want to encourage you to submit your amazing machine learning or deep learning use cases by **May 20.** Let me know in case you have any questions!

Happy Splunking,
Philipp

図-4 寄贈先Splunk Deep Learning Toolkit開発者メッセージ*4

*4 splunk.com、"Deep Learning Toolkit 3.1 - Examples for Prophet, Graphs, GPUs and DASK"（https://www.splunk.com/en_us/blog/machine-learning/deep-learning-toolkit-3-1-examples-for-prophet-graphs-gpus-and-dask.html）。

3.5 NLP(Natural Language Processing)を使ったテキストマイニング

NLPを使ったテキストマイニングでは、語彙間の関係性の分析や固有表現抽出で得られた情報を元に文章の全体像の把握や特徴抽出を行います。

Splunk Deep Learning ToolkitのNLPはDockerコンテナで稼働しているJupyterと連携して動作しており、アルゴリズムはPython Natural Language Processing libraryであるspaCyを用いて実装されています。

Entity	Entity_Count	Entity_Type	Entity_Type_Count
183万円	150	MONEY	42
1億円	96	MONEY	42
5月5日	96	DATE	55
92%	95	QUANTITY	108
日本	87	GPE	15
1万円	63	MONEY	42
9割	63	PERCENT	20
100%	56	QUANTITY	108
250万円	54	MONEY	42
100万円	52	MONEY	42
15分	49	TIME	16
1つ	43	QUANTITY	108
100人	42	QUANTITY	108
4000万円	41	MONEY	42
10分間	36	TIME	16
100%	34	PERCENT	20
火	33	DATE	55
11年	32	DATE	55
30万人	32	MONEY	42
第2267号	32	ORDINAL	10
800人	31	QUANTITY	108
橋本純樹	31	PERSON	45
3000万円	30	MONEY	42
92%	29	PERCENT	20
ワンクリックスキャル24/7完全無料公開中	28	PRODUCT	19
1割	25	PERCENT	20

表-1 2020年5月1日に定点で受信したスパム検体の固有表現抽出結果例

日本語テキストを処理可能にするため、Dockerコンテナイメージをカスタマイズし、spaCy 2.3.2へのアップグレードと追加された日本語モデルを含めた各国言語モデルの導入を行っています。

固有表現抽出のアルゴリズムはJupyter notebookで記述されているため、容易にカスタマイズが可能です。

表-1は定点で受信した2020年5月1日の1日分のスパム検体の本文データを独自拡張した固有表現抽出アルゴリズムで分析した結果になります。モデルはja_core_news_md(詳細はhttps://spacy.io/models/jaを参照)を使用しています。Entityが固有表現、Entity_Countがその固有表現の出現数、Entity_Typeが使用したモデルの中で定義されている属性分類、Entity_Type_Countがその属性分類の出現数を示しています。

人(PERSON)、お金(MONEY)、地名(GPE)、日付(DATE)や時間(TIME)数量(QUANTITY)などが抽出されています。プロダクト(PRODUCT)に該当する文字列が単語に分解されずに抽出されている点が注目されます。

この表ではEntity_Count数が大きい順にソートして出力していますが、Entity_Typeの箇所を見るとMONEYが上位を占めており、この日のスパムはお金に関する内容が記載されているものが多かったことが分かります。

表-2は同一の日のスパムを分析し、人名の属性を示すPERSONで絞り込んだ結果の抜粋です。人名が姓と名に分解されずに抽出されており、人名を解析軸として分析する場合に大きなメリットになります。大量のテキストデータを固有表現抽出により人名やプロダクト名で分類することができるため、稼働状況の分析やナレッジのデータベース化などに活用できそうです。

次に定点取得したスパム2020年の2月分と5月分の固有表現抽出結果でどのような差異が現れるか調べるためにそれぞれ上位15件の固有表現をグラフ化してみると、図-5と図-6のような結果となりました。

2020年2月ではまだコロナ禍の初期で海外旅行も行われていたことを反映しているのか、英語:LANGUAGEが最上位で相対的な比率でも突出して多い状況が分かりますが、緊急事態宣言後の5月では英語:LANGUAGEは大分ランキングを落として属性MONEYのものと入れ替わり、絶対数自体も大分増えていてスパム活動が活発化していることも分かります。

テキストデータ分析が難しい背景に、分類情報がなく解析軸が定まらないという点がありますが、このように固有表現抽出を用いることにより、テキストデータを固有表現の属性を使って分類可能になるので非常に大きな意義があります。

また、固有表現と属性分類の組み合わせ情報を利用することでテキストの概要パターンを識別可能になるため、テキストマイニングの可能性が大きく広がります。

Entity	Entity_Count	Entity_Type
橋本純樹	31	PERSON
佐々木千恵	22	PERSON
エリオット	17	PERSON
プロスペクト	17	PERSON
橋本	17	PERSON
佐々木	15	PERSON
トニー野中	9	PERSON
北条	9	PERSON
良彰	9	PERSON
アダム	8	PERSON
ロスチャイルド	8	PERSON
倉持	8	PERSON
サトー	7	PERSON
木村	7	PERSON
村岡	7	PERSON
よしあき	5	PERSON
ベール	5	PERSON
ザラ	4	PERSON
スキャルロジック	3	PERSON
たかはしよしあき	2	PERSON
カリスマ美人	1	PERSON
友宮真	1	PERSON
堀崎むつみ	1	PERSON
塚弥生	1	PERSON
大元大輝	1	PERSON

表-2 2020年5月1日に定点で受信したスパム検体で固有表現抽出を行い、PERSONで絞り込んだ結果の例

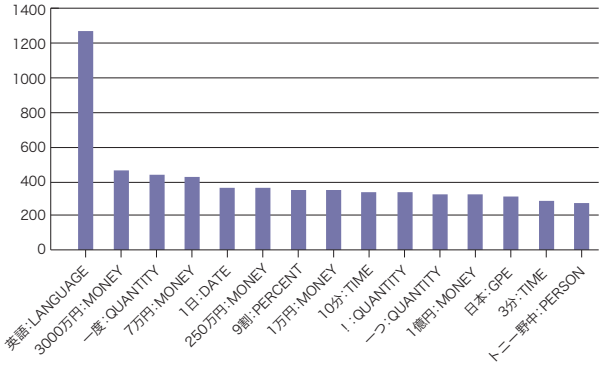


図-5 2020年2月の固有表現抽出結果上位15件のグラフ

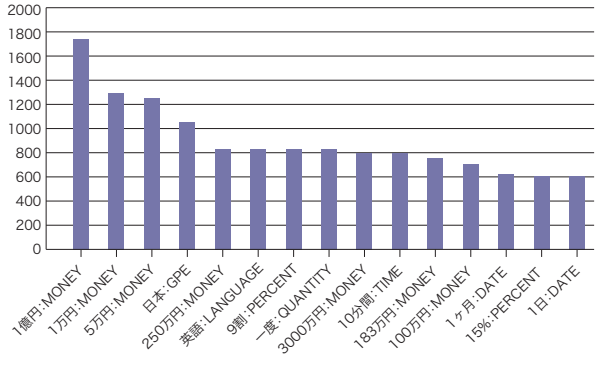


図-6 2020年5月の固有表現抽出結果上位15件のグラフ

3.6 テキストマイニングのビジネス活用

一般的にテキストマイングは様々なテキストデータをソースとして蓄積されるデータを元に、潜在ニーズの掘り起こしを目的として活用されています。

外部の音声テキスト変換APIなどを利用して音声データをテキストデータに変換し、ソースとすることも可能ですので、コールセンター業務などで蓄積される音声データを元にした顧客インサイト分析、業務上のナレッジ抽出などにも活用されています。テキストデータから事例のデータベースを構築し、似通ったパターンの事例を検索してマッチングするなどのユースケースがありますが、これらはニーズの掘り起こしだけでなく一内容の類似性による実績評価などに活用するケースがあります。

他社のサービス事例では、テキストチャット、音声チャットをチャットボットで一次受けを行い、それらのテキストデータを分析して必要に応じて人間による対応にエスカレーションさせるための仕組みの中で活用されています。例えばコールセンター業務の省人化によるコストダウンを目的としたサービスとして上手く建付けが行われている事例が見られます。

3.7 まとめ

従来大量のテキストデータの活用は難しくダークデータと化していましたが、現在では自然言語処理の精度向上により、テキストマイニングを幅広く活用することで有用な情報の掘り起こしが可能になってきています。

Splunk Deep Learning Toolkitのようにデータ蓄積からテキストの自然言語処理、モデル生成から、テキストマイニングまでシームレスに実行できる環境もあります。昨今注目されているテキストマイニングを始めてビジネスへ活用してみては如何でしょうか。

執筆者:
鈴木 徹(すずき とおる)
IIJ ネットワーククラウド本部アプリケーションサービス部xSPシステムサービス課シニアエンジニア。
GOJAS(日本Splunkユーザ会)運営メンバー。Splunkを活用してサービスに付加価値を生み出す活動を行う。

Information

IIJ 技術情報発信コンテンツの紹介

■ IIJ Engineers Blog 最新トピック

開発・運用の現場エンジニアが執筆するIIJ公式ブログ「IIJ Engineers Blog」では、旬なテーマからエンジニアの趣味・面白ネタまで、さまざまなトピックの記事を掲載しています。新入社員が配属される時期には、「新人エンジニアにおススメする技術書」をテーマに、IIJのエンジニアが選んだ技術書を紹介しました。

■ 最近の掲載記事（https://eng-blog.iij.ad.jp/）

- The Internet Health Report
- 自宅にKubernetesクラスター『おうちKubernetes』を作ってみた
- Pure Python Tracepath
- 迷惑メールの量が急増中！ 2020/1Q 緊急レポート

■ シリーズ：新人エンジニアにオススメする技術書

- まとめて紹介！新人エンジニアにオススメする技術書（https://bit.ly/3i1fvEH）
  1.『体系的に学ぶ 安全なWebアプリケーションの作り方 第2版 脆弱性が生まれる原理と対策の実践』
  2.『リーダブルコード ―より良いコードを書くためのシンプルで実践的なテクニック』
  3.『レジデント初期研修用資料 医療とコミュニケーションについて』
  4.『ライト、ついてますか―問題発見の人間学』
  5.『改訂新版 コンピュータの名著・古典 100冊』
  6.『アカマイ 知られざるインターネットの巨人』
  7.『使える力が身に付く DNSがよくわかる教科書』
  8.『小悪魔女子大生のサーバエンジニア日記～インターネットやサーバのしくみが楽しくわかる』
- 『人月の神話』
- 『プログラミング作法』
- 『テスト駆動開発』
- 『UNIXという考え方』

■ IIJ Technical NIGHT vol.9 開催報告

去る9月11日(金)、コロナ禍で3月から延期していた「IIJ Technical NIGHT vol.9」を完全オンラインの形式で開催しました。

■ テーマ「セキュリティアナリストのお仕事――分析、AI、ツール開発」

IIJのSOC(Security Operation Center)での取り組みを紹介。サイバー攻撃に対応するため24時間ネットワークを監視するSOC。しかし、武器がなくては日々のセキュリティ・インシデントに立ち向かう事はできません。IIJでは、SOCのアナリストを支援する武器として、一般的に流通している情報やツールのほかに、社内のエンジニアが独自の方法で分析したり、ツールの開発を行っています。今回はその中から、「インシデントを発見するためのインテリジェンス構築」と「インシデント分析のためのツール開発」における3つの取り組みを、実際にSOCで働いている中の人が紹介しました。

【プログラム】

Session1:あ！ やせいのEmotetがあらわれた！ ～ IIJ C-SOCサービスの分析ルールについて ～
Session2:セキュリティとAIと私
Session3:インシデント調査システムが内製すぎる件 ～ CHAGEのご紹介 ～

当日の資料は以下のブログ記事で紹介しています(https://eng-blog.iij.ad.jp/archives/6453)。

■ IIJ 公式Twitterアカウント@IIJ_ITS

なお、今回ご紹介した「IIJ Engineers Blog」、「IIJ Technical NIGHT」で更新があった際には、IIJ公式Twitterアカウント@IIJ_ITSでお知らせしています。ご興味のある方はぜひフォローしてください。
@IIJ_ITS　https://twitter.com/IIJ_ITS

September 2020 Vol.48

IIJ
Internet Initiative Japan

株式会社インターネットイニシアティブ(IIJ)について

IIJは、1992年、インターネットの研究開発活動に関わっていた技術者が中心となり、日本でインターネットを本格的に普及させようという構想を持って設立されました。
現在は、国内最大級のインターネットバックボーンを運用し、インターネットの基盤を担うと共に、官公庁や金融機関をはじめとしたハイエンドのビジネスユーザに、インターネット接続やシステムインテグレーション、アウトソーシングサービスなど、高品質なシステム環境をトータルに提供しています。
また、サービス開発やインターネットバックボーンの運用を通して蓄積した知見を積極的に発信し、社会基盤としてのインターネットの発展に尽力しています。

本冊子の情報は2020年9月時点のものです。

IIJ-MKTG019-0048

株式会社インターネットイニシアティブ
〒102−0071 東京都千代田区富士見2-10-2　飯田橋グラン・ブルーム
E-mail：info@iij.ad.jp　URL：https://www.iij.ad.jp