基礎研レポート

テキストマイニングを用いて景況を把握する

金融研究部 研究員　水野　友理那
(03)3512-1856　y-mizuno@nli-research.co.jp

1――活用が進むテキストマイニング

ここのところ、国内において、テキストマイニングを利用した商品やサービスを目にすることが多くなった(図表１)。テキストマイニングとは、文章からなるデジタルデータ（以下、テキスト情報）の分析を行うことである。テキスト情報を名詞や動詞などの単語に分け、単語の出現頻度や出現パターンの分析などにより、目的とする情報を取り出す方法だ。コンピューター等の力を借りることで、人が読み込むことが不可能なほど膨大なテキスト情報を分析することが可能となっている。

図表１：テキストマイニングが利用される商品やサービスの例

公表年月	企業	内容
2017/2	三菱UFJ国際投信	ニュースや決算報告書などの分析結果を運用戦略に利用するアクティブファンドの販売
2017/9	アセットマネジメントOne	ニュースや決算報告書などの分析結果を運用戦略に利用するアクティブファンドの販売
2018/3	ISIDほか	企業・商品への評価や市場動向への調査を目的とするソーシャルメディア分析ツールの提供
2018/4	NTTデータ、角川アスキー総合研究所	映像などのコンテンツに関して投稿されるツイートデータを分析した視聴者分析レポートの提供
2018/10	日立製作所ほか	企業・商品への評価や市場動向への調査を目的とするソーシャルメディア分析ツールの提供
2018/10	NEC、ダンデライオン・チョコレート	過去の新聞記事から当時のムードを味わえるチョコレートの開発

（参考）新聞記事等から筆者作成

1｜多情報社会では、情報の集約に価値がある

近年、手軽に膨大な情報を得られるようになった。日本全国で発行される新聞はおよそ 100 紙[1]あり、一般雑誌は週刊・月刊合わせて 1,000 誌[2]程もある。さらに、インターネットを介して発信される情報も増えた。総務省「メディア・ソフトの制作及び流通の実態に関する調査」によると、インターネットオリジナルのテキスト情報の量[3]は、2011 年は 982 億頁（B5 版換算）だったが、2016 年には 1301 億頁と 30%も増加した。

実際に、全てのテキスト情報を読む人はいない。従来から、人はそれぞれ工夫を凝らし、効率的な情報収集を試みてきた。新聞を読む際は先に見出しに目を通すなど、各人が必要な情報を選択している。最近では、自分の関心に沿うニュースを日々届けてくれるキュレーションメディアなどのサービスや、ニュースを要約するサイトもある。情報が多い現代では、必要な情報の選別や集約は非常に重要である。

2｜膨大な文章の分析を可能にするテキストマイニング

しかし、各人の判断で選別すると、自分が興味を持つ断片的な情報のみを集めることとなり、網羅性に欠ける恐れがある。その結果として、特定の意見への偏重などが危惧される。そこで、網羅性を確保しつつ効率的に情報を集約するために、テキストマイニングが使われてきている。

例えば、図表１に示した２社の例では、アクティブファンドの運用戦略にテキストマイニングを利用している。通常、アクティブファンドの多くは、投資対象銘柄を選択する際に、アナリストによる業績予想を参考にする。アナリストが分析する対象は、株価・財務データなどの定量データに加え、有価証券報告書やニュースなどのテキスト情報など、あらゆる情報だ。定量データを分析することは、コンピューターの性能が高くなった現代ではさほど大きな負荷にならないが、テキスト情報の分析は容易ではない。東証一部上場銘柄だけでも 2,000 企業程度もある。各企業が公表するテキスト情報の全てをアナリストが読んで評価するには限界があることから、全ての上場企業を分析することはできない。一方、テキストマイニングを利用すれば、テキスト情報の分析負荷が軽減される。分析可能な企業を増やすことによって、収益獲得機会の拡大に繋がる。このように、膨大なテキスト情報を取り扱う場合は、テキストマイニングが有効であると言えるだろう。

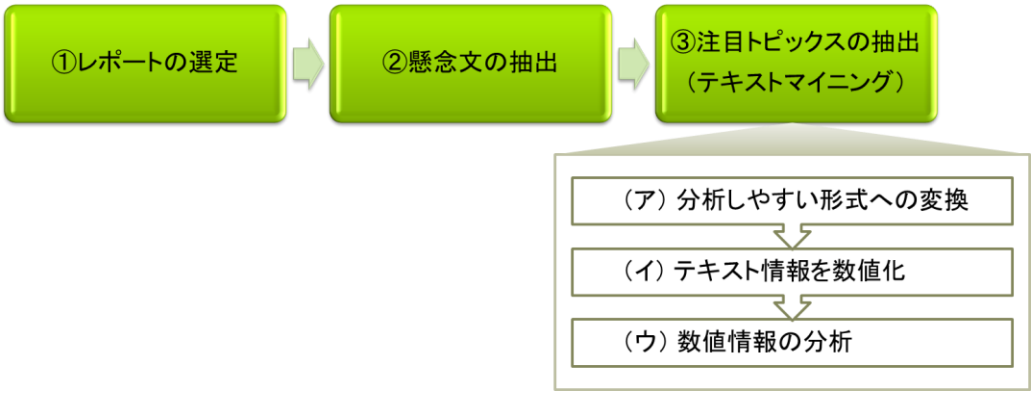
この他、マーケティング分析・商品開発の分野でも、テキストマイニングは利用されている。これまでにも、自由記述形式のアンケート調査を分析する際に使われていたが、ソーシャルメディアの普及により注目度はより高まった。膨大な情報が発信され続けるソーシャルメディアには、企業・商品に対する消費者のリアルタイムな評価が含まれる。アンケートを実施しなくとも、ソーシャルメディアから消費者のニーズなどを効率的に得ることができる。加えて、リアルタイムであることも利点だ。

2――エコノミストが注目する景況下方要因とは何か

上記のように、テキストマイニングは、業務効率の向上や付加価値の創出に活用されている。そこで、テキストマイニングを用いて、数多く公表されるテキスト情報の中から、実際に注目トピックスを抽出してみたい。本稿では、注目トピックスを、景況見通しの下方要因として多くの有識者が注目していたと推測される単語とし、具体的な分析の手順、抽出結果を示す。最後に、テキストマイニングを利用する際の留意点について、分析者の立場から意見を述べたい。

分析手法の検討にあたり、複数のテキスト情報から情報を取り出す際に、人が行う通常の工程を再現することに重きを置いた（図表２）。初めに、エコノミスト等が景況の見通しについて言及するレポートを選定する。その中から、景況への下方要因が記述された文（以下、懸念文）を抽出する。懸念文の中から、他の月より注目を多く集めた単語を各月の注目トピックスとして抽出する。

図表２：分析手法の手順



1｜分析手法

①　レポートの選定

レポートの選定にあたり、日本経済の見通しに言及されていること、定期刊行されていること、表現が明瞭であること、公表主体の網羅性（異なる金融系列や業態）を重視した。金融商品の販売促進を目的としたものや、ファイルが保護されており文章を抽出できないものなどは排除した。その結果、６団体から、1998 年２月～2018 年８月までのレポート 826 本を選定した。

②　懸念文の抽出

上記①で選定したレポートの全文から、懸念文を抽出した。将来の景況下方要因について言及している文を懸念文とし、２つの条件を設定した。１つ目は、「懸念を表す単語」が入る文であることだ。「懸念を表す単語」として、「懸念」「リスク」「可能性」「不安」など、計 28 単語を筆者の主観により選定した。２つ目の条件は、過去を表す助動詞が、文の後半に混じっていないこととした。結果として、懸念文総数は、11,524 文となった。

③　注目トピックスの抽出（テキストマイニング）

懸念文を分析し、以下（ア）～（ウ）の手順で注目トピックスを抽出した。注目トピックスは具体的に、他の月と比較して「頻出度」が高い単語を想定している。

(ア) 分析しやすい形式（単語の羅列）への変換

テキストマイニングでは、単語の出現傾向を分析する。まず、文章中の単語の品詞を識別し、単語ごとに区切る（文章の分割）。ただし、「文章の分割」だけでは、同じ事柄を指し示す単語や書式など、書き手の表現方法の違いを調整できず、注目度の評価（「頻出度」のカウント）に支障をきたすため、「単語の統一」「助詞・記号の削除」なども行った（具体的には図表３に示す）。このように、文章を「単語の羅列」に変換した。変換前後の具体例は、以下の通りだ。

変換前：トランプ大統領の発言を受け、為替市場は
変換後：米国大統領 / 発言/ 受ける / 為替市場

図表３：分析しやすい形式への変換

文章の分割	• 品詞の識別 • 単語ごとに区切りを入れる
単語の統一	• 単語を原形に変換 • 同義の単語を統一 • 書式の統一（カナは半角に、アルファベットは小文字半角に統一するなど）
助詞・記号の削除	• 「が」「は」「に」「、」などの助詞・記号の削除

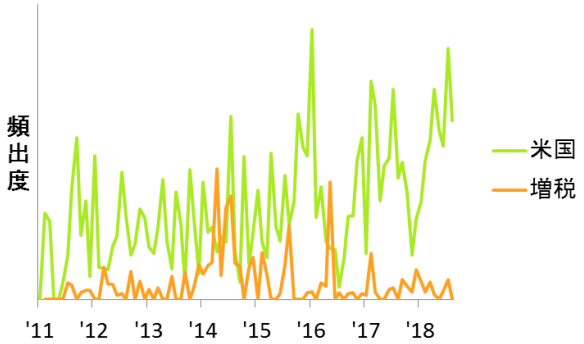
(イ) テキスト情報を数値化

次に、注目トピックスとして抽出する単語を絞り、「単語の羅列」をもとに、抽出した単語に対して「頻出度」を計算する。懸念文に含まれる単語は、名詞と動詞に限定しても 5,000 単語以上あり、下方要因に結びつけにくい単語の抽出やシステム負荷の増大などの不都合が生じるため、頻出度を算出する単語は 300 単語ほどに絞った。頻出度は、懸念文として抽出される文章数に対する、各単語が使用される数の割合とし、各月において相対的に多く使用される単語ほど高くなるようにした。各月の懸念文の数に月ごとの違いがあり、同月内の他の単語と、また、他の月の同単語とも公平に比較可能となるように補正する必要があるからだ。

(ウ) 数値情報の分析

頻出度の平均値やそのばらつきを参考に、他の月に比べて頻出度が異常に高い単語を注目トピックスとして抽出した。各月を代表する単語を人が判別する際、過去の記事で言及されていなかった単語ほど、特別視されると考えられるからだ。例えば、「米国」のように毎月のように使われるような単語は注目トピックスとして認識されにくい（図表４）。一方、「増税」は、2014 年の消費税を８％に引き上げる前後や、10%引き上げの延期が決定する前の 2015・2016 年など、特定の時期にのみ頻出度が高く、注目トピックスとして認識される可能性が高い。

図表４：頻出度の推移



2｜分析結果

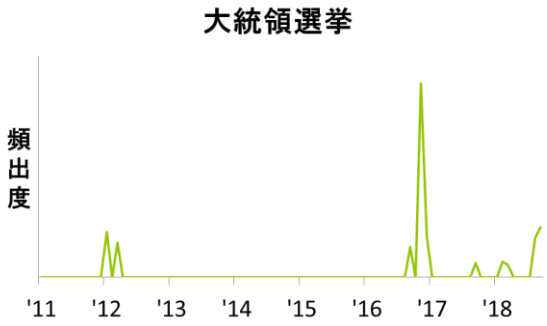
上記の分析手法の結果、筆者の感覚として妥当な単語が各月における注目トピックスに抽出された。抽出された注目トピックの例として、図表５に 2016 年 11 月の結果を示す。「大統領選挙」が突出した注目トピックスとして抽出された。同月は、トランプ大統領が当選確実となった月であり、過去の事実と違和感のない妥当な結果が得られた。

「大統領選挙」の頻出度は 2012 年前後にも抽出されたが（図表６）、2012 年はオバマ大統領２期選にあたる。2016 年 11 月の「大統領選挙」の頻出度は、2012 年の 4 倍程度を示しており、トランプ大統領への関心が異常に高かったことがうかがわれる。

図表５：2016 年 11 月の注目トピックス

注目トピックス（2016年11月）
大統領選挙
大統領
増益
減額
リスクオフ

図表６：「大統領選挙」の頻出度の推移



3――留意点と今後について

テキストマイニングは、膨大なテキスト情報を分析する際に有用であると述べた。しかしながら、テキストマイニングを利用する際の留意点を３点挙げる。まず、分析対象とするテキスト情報を適切な量だけ入手できるかどうかだ。テキストマイニングでは単語を統計的に処理するため、適切な分析には相応の量が必要である。２つ目は、意図する目的に対して、膨大なテキスト情報を分析することが効率的かどうかだ。例えば、辞書に収録されている単語の意味を知りたい場合は、わざわざ膨大なテキスト情報を分析するまでもない。３つ目は、目的に則した分析手法の検討に、相当の時間がかかる点だ。分析結果が何を示すかは、「分析対象とする文章の選定」と「数値情報の分析方法」に依存するため、意図する目的に合致する結果が得られるか、検証を重ねる必要がある。これら多くの労力をかけても、それに値する効果が見込めれば、テキストマイニングは目的を達成するための有効な手段となる。

最後に、上記の観点から、本稿の分析手法における今後に向けた改善点を述べる。アナリストが言及する景況への下方要因の抽出という目的に照らし、「分析対象とする文章の選定」をより適切なものにするための改善点は２点ある。１点目は、分析対象とする専門家のレポートの拡張である。異なる書き手の文章を多く分析するほど、下方要因を網羅的に分析でき、「総意」としての精度が高まることが期待される。２点目は、懸念文抽出方法の高度化である。その手段として、今回は筆者の主観で選んだ「懸念を表す単語」のより合理的な選定などが挙げられる。また、「数値情報の分析方法」をより高度化することで、分析結果が示す意味は変わる。例えば、単語同士の関係性を定量的に分析することで、景況の下方要因となりうる事象の因果関係の構造化が可能となるかもしれない。今後、より高度な分析を試みていきたい。

[1] 日本新聞協会会員社により発行される新聞の総数(2018 年 11 月時点)
[2] 日本雑誌協会会員社により発行される雑誌の総数(2018 年 11 月時点)
[3] メールマガジン、広告、ブログ、SNS などの合計値