令和３年２月４日

景気ウォッチャー調査を用いたテキスト分析の方法について（第１回）

茨城県政策企画部統計課　企画分析グループ

１　はじめに

人々の景況感を掴むことを目的に、茨城県統計課（以下「当課」という。）が実施している「茨城県景気ウォッチャー調査[1]」（以下「本調査」という。）は、約 300 名の調査客体の景気判断理由をテキストベースで得られるという特徴がある。今回は、このような数値化できないデータを視覚的に分かりやすくして俯瞰できるようにするテキスト分析の方法として、単語間の共起（ある２つの語が同じ文章中に出現すること。）関係を可視化する「共起ネットワーク図」を中心に紹介し、調査客体の景気判断理由を考察する。

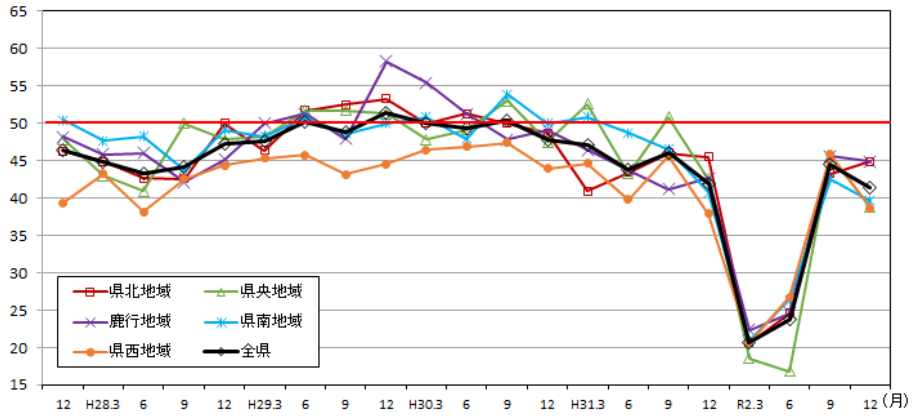
なお、分析にあたってはプログラムを書かなくても共起ネットワーク図を作成することができる「KH Coder」[2]を用いる。また、今回の分析に用いるデータは当課公式サイト[3]で公表している。

２　調査結果からの推測

景気の現状判断 DI（調査時点の景況感を示す指標。以下「DI」という。）の推移を見ると、令和２年以降の動きに特徴があることがわかる（図１）。令和２年３月調査では、DI 値は大きく低下しており、リーマン・ショック期の DI 値[4]と同程度にまで落ち込んだ。その半年後の９月調査では、未だ DI 値が 50 を下回っているものの、大きく低下する前の令和元年 12 月調査と同程度の DI 値まで上昇した。

令和２年３月頃は、国内で新型コロナウイルスの感染が拡大し始めた頃である。令和２年３月調査の調査客体の景気判断理由のコメントでも、新型コロナウイルスに関するものが多く見られたことから、新型コロナウイルスが DI 値の落ち込みに繋がったと推測できる。

【図１】　景気の現状判断DIの推移（地域別）



[1] 調査概要：https://www.pref.ibaraki.jp/kikaku/tokei/fukyu/tokei/betsu/bukka/watch/chogai.html

[2] テキストデータを統計的に分析するためのフリーソフトウェア。（http://khcoder.net よりダウンロード可能）

[3] https://www.pref.ibaraki.jp/kikaku/tokei/fukyu/tokei/betsu/bukka/watch/bunseki/index.html

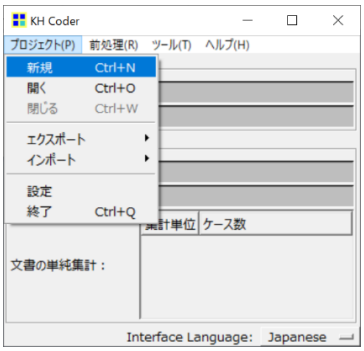
[4] リーマン・ショック期は DI 値が大きく低下し、平成 21 年３月には本調査開始以来過去最低の 18.5 となった。

３　分析データの読み込みと前処理

「KH Coder」を使って、景気判断理由に新型コロナウイルスに関するコメントがどのくらい多いか可視化してみよう。

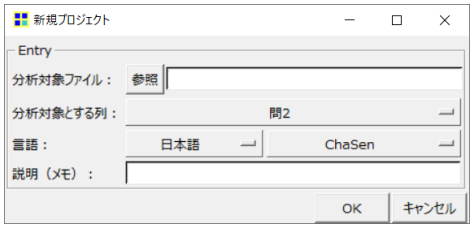
（１）新規プロジェクトの作成

KH Coder を起動し、[プロジェクト(P) － 新規]から新規プロジェクトを作成する。



（２）分析対象ファイルの読み込み

「参照」ボタンから、分析対象のデータファイルとして令和２年３月調査の景気判断理由のテキストデータをまとめた Excel ファイル「2020.03 茨城県景気ウォッチャーコメント.xlsx」を選択して読み込む。

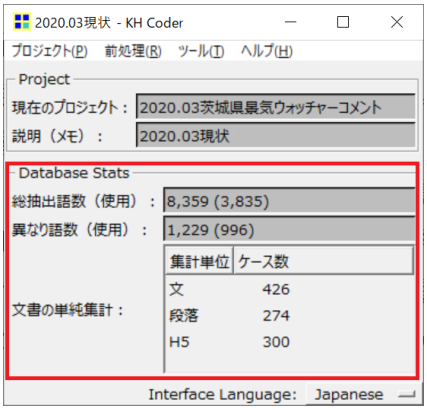


この際、分析対象とする列を指定する。ここでは、現状の景気判断理由が書かれている「問２」列を選択する。

（３）前処理の設定・実行

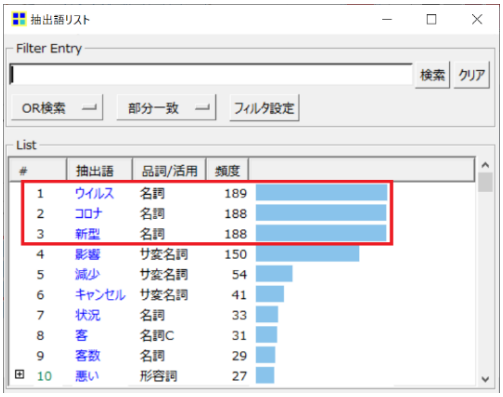
分析に先立っては、「形態素解析」と呼ばれる前処理が必要である。この前処理によって、分析対象のテキストデータから語が切り出され、頻出語はどのようなものがあるかなどの確認ができるようになる。前処理には、形態素解析のための辞書が必要であり、KH Coder では、「MeCab」か「ChaSen」の２種類が用意されているが、ここでは「MeCab」を選択し、「OK」ボタンを押す。

次に、[前処理(R) － 前処理の実行]から前処理を実行する。前処理が完了すると、「Database Stats」内に前処理の結果が表示される。



（４）抽出語リストの確認

［ツール(T) － 抽出語 － 抽出語リスト]から、頻出上位の語を確認する。すると、「ウイルス」、「コロナ」、「新型」などの語が頻出していることが分かる。

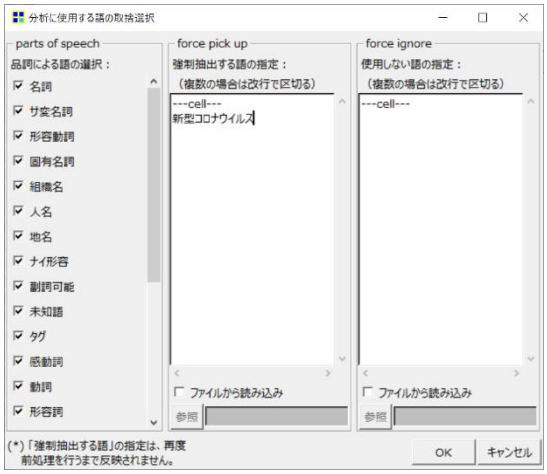


ここで、「新型コロナウイルス」といった最近になって使われ始めたような語は、先述した辞書には登録されていないため、「新型コロナウイルス」１語では切り出されない。

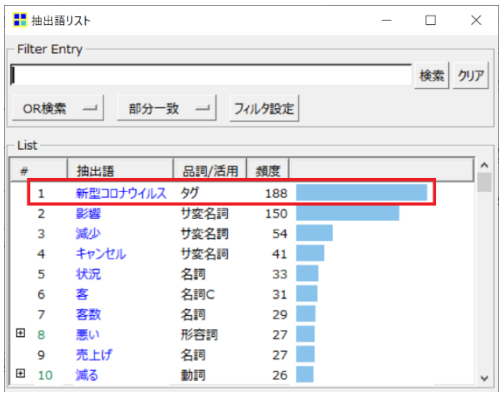
このような場合には、「新型コロナウイルス」を１語で強制的に抽出させる設定を加える。

（５）強制抽出する語の指定

［前処理(R) ― 語の取捨選択］から、強制的に抽出させたい語を左枠「force pick up」内に入力する。



強制抽出する語の指定が完了したら、［前処理(R) ― 前処理の実行］から再度前処理を実行する。前処理の完了後、抽出語リストに「新型コロナウイルス」の語が正しく強制抽出されていることを確認する。

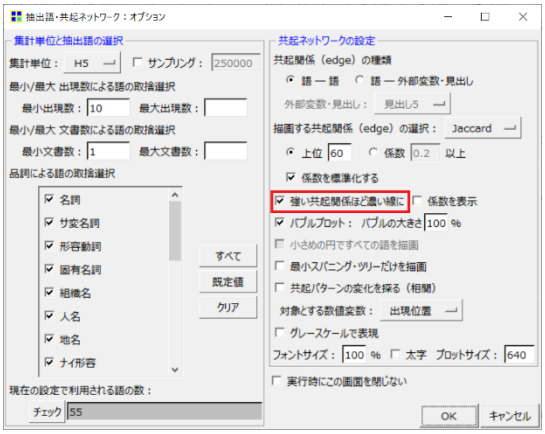


４　語と語の共起関係

前処理が完了すると、共起ネットワーク図の作成などの分析ができるようになる。

［ツール(T) ― 抽出語 ― 共起ネットワーク］から、共起ネットワーク図の設定を行う。

ここでは、「品詞による語の取捨選択」などの項目は既定値のままで行う。ただし、共起度の強弱を可視化するため、「強い共起関係ほど濃い線に」にはチェックを入れることを推奨する[5]。



以上の設定で共起ネットワーク図を作成した結果は【図２】[6]のとおり。この図を見ると、「新型コロナウイルス」という語に、「キャンセル」や「減少」といった語が共起している。「減少」には「客数」も共起していることから、新型コロナウイルスの影響により、キャンセルや客数の減少が起こっていることが読み取れる。実際のテキストデータを覗いても、そのようなコメントが複数見られた。

---

[5] 共起の強弱をはかる指標として Jaccard 係数などが用いられる。ここでは、既定の設定である Jaccard 係数により共起の強弱を計算させている。

[6] この図では比較的強い共起関係にある８つのグループが８色で色分けされているが、色そのものには意味はない。

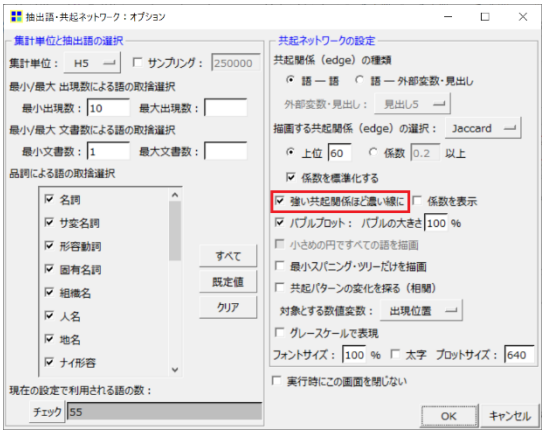
＜調査客体のコメント＞（一部抜粋）

・ゴルフ場（県央地域）

例年３月は来場客数がピークとなるが、新型コロナウイルスの影響から、キャンセル数も多く、前年比で 20%以上の減少が予想される。

・旅行代理店（鹿行地域）

新型コロナウイルスの影響により、予約のキャンセルが相次ぎ、客数が減った。



５　語と景気評価の共起関係

「４　単語間の共起関係」では、語ー語での共起関係を図にしたが、今度は語ー外部変数・見出しでの共起関係を見ていく。

「外部変数・見出し」で、令和２年３月調査の「問１」（調査月における景況感についての評価を問うもの）を選択すると、問１の各回答（「1：良くなっている」「2：やや良くなっている」「3：変わらない」「4：やや悪くなっている」「5：悪くなっている」）と共起している語の様子を見ることができる。

例えば、ある語「word」が「4：やや悪くなっている」や「5：悪くなっている」と共起していれば、調査月における景況感で、４や５のマイナスな評価をしている人のコメントに「word」という語が多く見受けられると言うことができ、「word」が DI 値の低下要因になっている可能性が高い。

では、［ツール(T)　ー　抽出語　ー　共起ネットワーク］から、共起ネットワーク図の設定を再度行ってみる。

「共起関係（edge）の種類」を「語ー外部変数・見出し」に変更し、「外部変数・見出し」には「問１」を選択する。

作図した結果は、【図３】[7]のとおり。赤い四角は「外部変数・見出し」に指定した問１の各回答番号で、四角内の数字が大きいほど調査月の景況感に対する評価が低い。この図から、次の２点に注目したい。

① 調査月の景況感の評価が「３：変わらない」以下の評価において「新型コロナウイルス」と共起しており、特に、４や５の評価で共起度が高い。

② 「４：やや悪くなっている」や「５：悪くなっている」といった低い評価において「キャンセル」と共起している。

これらの結果から、新型コロナウイルスについては、「２　調査結果からの推測」で述べた「新型コロナウイルスが DI 値の落ち込みに繋がった」とする推測のとおりであることが見て取れる。また、「キャンセル」という語は出現頻度が比較的高かった（41件）ものの、４や５といった低い評価にのみ共起関係が見られたという点から、「キャンセル」が調査客体の景況感をより下向きにさせ、令和２年３月調査の DI 値の下げ幅を

[7] この図では語が３色で色分けされているが、その色分け規則は語から出ている線の本数によっている。

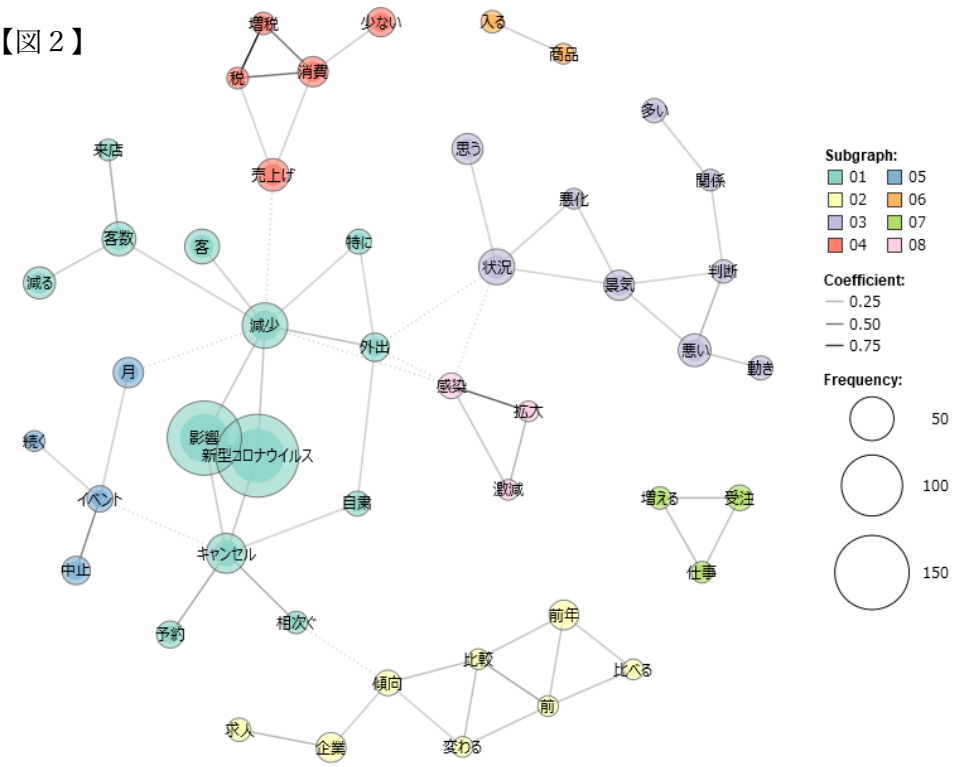
大きくさせたことが考えられる。

６　まとめ

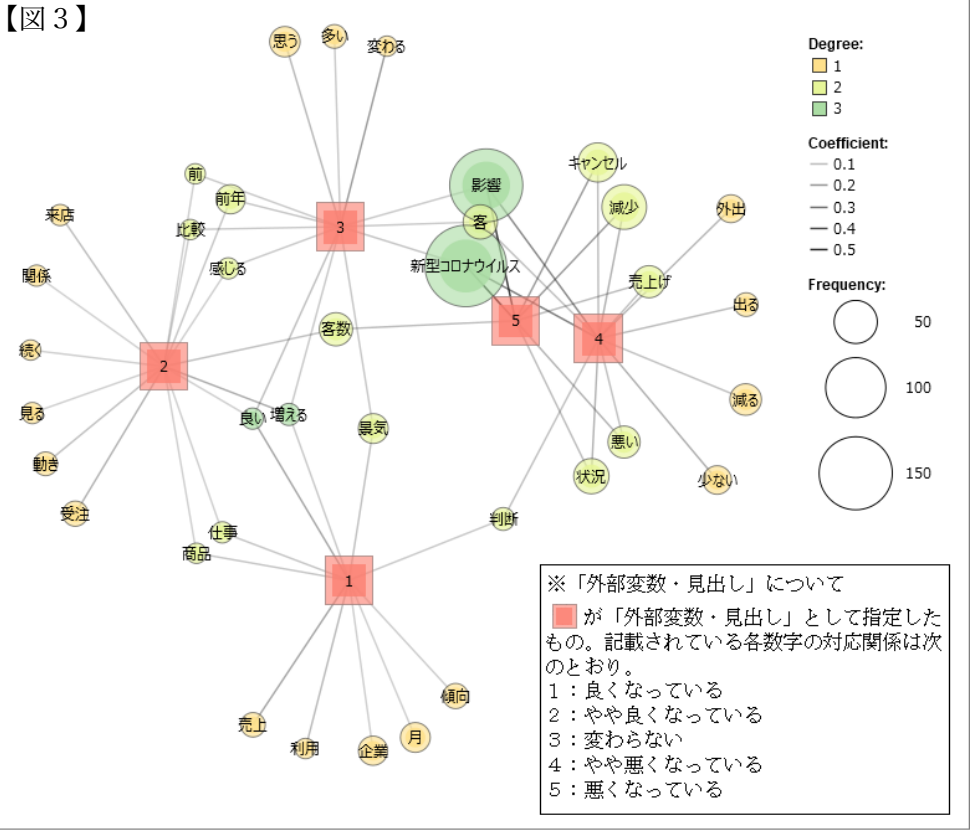
「４　語と語の共起関係」と「５　語と景気評価の共起関係」の２つの側面から共起ネットワーク図を作成した結果、新型コロナウイルスが景況感に悪影響を及ぼし、特に、キャンセルやそれに伴う客数の減少が DI 値の大幅低下を後押ししたという示唆が得られた。

以上のように、テキストデータを視覚的に分かりやすくして俯瞰するには、今回紹介した方法が有効である。とりわけ、本調査はテキストの量が多く、全て読むには時間を要するため、内容や特徴を大まかに把握したい場合には特に有効となる。

【図２】



【図３】



※「外部変数・見出し」について

■が「外部変数・見出し」として指定したもの。記載されている各数字の対応関係は次のとおり。

１：良くなっている
２：やや良くなっている
３：変わらない
４：やや悪くなっている
５：悪くなっている