特許マップ作成におけるデータマイニング及びテキストマイニングの活用

新電元工業株式会社　知的財産部
阿河正明

１．はじめに

特許情報については、データ化されており、容易に収集することができる。また、収集した情報をデータから抽出し、解析（マイニング）することで、今後の開発テーマを発見することができる場合もある。

最近は、特許情報をデータ化し、又はテキスト化し、マイニングするソフトウェアも充実しており、このようなソフトウェアを活用することで、人間の手では解析できない膨大な量のデータを瞬時として解析することができる。

但し、どのような場合にデータマイニングやテキストマイニングを活用したらよいか、不明なところが多い。

そこで、どのような場合に、データマイニングやテキストマイニングを活用すべきかを報告する。

２．データマイニングについて

２．１　データマイニングとは

データマイニング（英語: Data mining）とは、統計学、パターン認識、人工知能等のデータ解析の技法を大量のデータに網羅的に適用することで知識を取り出す技術のことである(「Wikipedia」より)。DM と略して呼ばれる事もある。

２．２　データマイニングと特許マップとの関係

特許文献は、単なる文章のみで構成されておらず、多くのデータが存在する。なお、後述するテキストマイニングと区別するため、この章では、テキストデータについては除く。データマイニングを行う上のデータとして、例えば、特許文献から IPC（国際特許分類）、FI（File Index・IPC を細分化した日本国特許庁独自の特許文献の分類）、Ｆタームなどの分類データや、登録日、公開日、出願日などの日付データのみならず、特許出願人（特許権者）、発明者などの名前データなどが、活用できる。これらのデータを組み合わせることにより、特許マップを作成することができる。例えば、ある分類の出願人データと日付データとを組み合わせれば、ある出願人がどの時期にその分類について力を入れたのかを解析することができる（図１参照）。また、ある出願人データとより細かい分類データとを組み合わせれば、ある出願人がどの分野に力を入れているかを解析することができる（図２参照）。

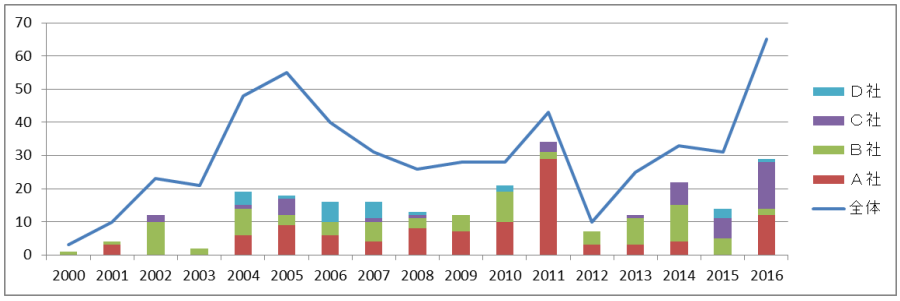

図１:燃料電池ステーション（Ｆターム：5H127FF20）の年度別出願件数

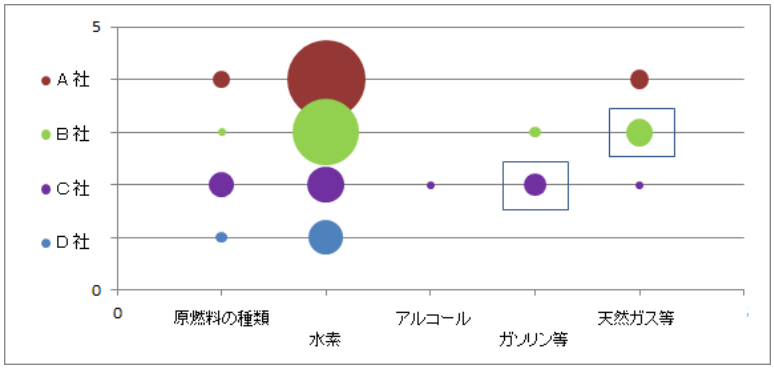

図２：燃料電池ステーション（Ｆターム：5H127FF20）の燃料の種類別出願件数

４社は業界地位として、Ａ社はリーダー企業、Ｂ社はチャレンジャー企業、Ｃ社はニッチャー企業、Ｄ社はフォロワー企業に該当する。Ａ社は圧倒的に最も実用性の可能性が高い「水素」を燃料とした燃料電池に力を入れている。これに対して、Ｂ社は「水素」のみならず、Ａ社と差別化を図るため「天然ガス等」を燃料とした燃料電池にも力を入れていることが分かる。Ｃ社は４社のうち唯一自動車製造会社ではなく、燃料供給会社であり、Ａ社、Ｂ社と差別化を図るため、Ｃ社で供給している「ガソリン等」を燃料とした燃料電池にも力を入れていることが分かる。Ｄ社はフォロワー企業らしくリーダー企業のＡ社と傾向が似ている。

３．テキストマイニングについて

３．１　テキストマイニングとは

テキストマイニング（英語: Text mining）とは、文字列を対象としたデータマイニングのことである。通常の文章からなるデータを単語や文節で区切り、それらの出現の頻度や共出現の相関、出現傾向、時系列などを解析することで有用な情報を取り出す、テキストデータの分析方法をいう（「Wikipedia」より）。

３．２　「分かち書き」と「係り受け」

テキストマイニングにおいては、「分かち書き」作業と、「係り受け」作業が重要である。

「分かち書き」作業とは、文章からなるデータを単語や文節で区切る作業をいう。この「分かち書き」作業を行うことにより、単語頻度分析を行うことができる。

また、「係り受け」作業とは、文法的知識を用いて、分かち書きした単語や文節の構文分析を行う作業をいう。この「係り受け」作業を行うことにより、係り受け頻度分析を行うことができる。

３．３　テキストマイニングと特許マップとの関係

３．３．１　テキストマイニングに最適な文章の選択

特許文献には、特許請求の範囲、明細書、要約書などに多くの文章が記載されている。これらは、テキストマイニングにおけるデータになり得る。これらのうち、明細書や要約書に用いられる文章は平素な文章で記載されていることが多い。そのため、「係り受け」作業が容易であり、テキストマイニングに適している。一方、特許請求の範囲の記載方式として、ジェプソン方式、書き流し方式、要件列挙方式があるが、このうち、最近最もポピュラーな要件列挙方式は、一または複数の発明を箇条書きにした形式をとり、さらに名詞句で記載する特殊な文章形式で記載されている。このような文章形式は、修飾語が多く、「係り受け」作業が非常に難しい。また、「係り受け」作業は、テキストマ

イニングができるソフトウェアを用いて行うが、この「係り受け」作業をするのに時間がかかる上に、平素な文章で記載される場合と異なり、修飾関係が複雑であるため、「係り受け」を誤ることが多い。そのため、特許請求の範囲は、テキストマイニングにはあまり適さない。

上述の通り、テキストマイニングには、明細書か要約書が適している。

但し、明細書は、幾つもの段落から構成されており、文章の量が制限されていない。そのため、後述する単語頻度や係り受け頻度は単語数の多い特許文献の影響を受け易く、信憑性に欠ける恐れがある。また、文章の量が膨大なため、解析に時間がかかるという欠点もある。

一方、要約書は、４００字以内の限定があり、明細書に比べると、単語頻度数の差が小さく、信憑性が高い。さらに、発明の簡単な構成と課題が記載されており、こちらが想定したい単語がヒットする可能性が高い。また、上述の通り、「係り受け」作業が比較的容易の上、文字数が少ないため、特許件数が多くなっても分析に比較的時間がかからないという利点もある。

従って、特許分析を行う際にテキストマイニングに用いる文章としては要約書を用いることが最適である。

３．３．２　テキストマイニングの前処理

特許文章をテキストマイニングするためには、「分かち書き」作業をする前に、編集作業が必要である。これを前処理という。例えば、要約書は、概ね【課題】【解決手段】【選択図】の構成からなる。これらの文言を残しておくと、単語頻度分析の上位にヒットする。そのため、これらの言語は最初から除外すべきである。また、【選択図】の部分は全て削除しないと、例えば、「図１」が単語頻度分析の上位にヒットする。これらを前もって削除しておく。なお、ここからテキストマイニングを Text Mining Studio を使用して行う。

３．３．３　「分かち書き」と単語頻度解析

前処理が終了したら、「分かち書き」を行う。「分かち書き」を行うことで、品詞の出現回数が分かる（図３参照）。特許文章の特徴は、主語になる名詞が圧倒的に多く、続いて、「する」を加えることにより動詞になり得る名詞（以下「サ変可能名詞」という。）、動詞が多い。一般的な文章では形容詞や形容動詞も名詞や動詞に並んで多いが、特許文章においては圧倒的に少ない。このことから、特許文章において、良いイメージで語られることば、悪いイメージで語られることばを抽出する評判抽出は難しいと言われている。一方、図３で示す通り、特許文章は、名詞、サ変可能名詞、動詞の割合が極端に高いため、後述する係り受け頻度解析は、名詞、サ変可能名詞、動詞の３品詞に絞り込むことができる。

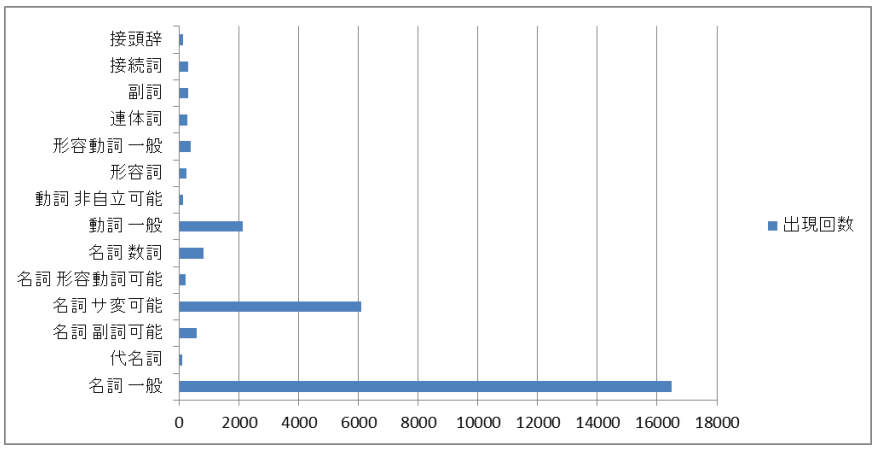

図３：燃料電池ステーション（Ｆターム：5H127FF20）に該当する特許文献のうち要約書で出現する品詞の出現回数を示す図

続いて、前処理をしても、単語頻度解析をするのに不要な言語が存在する。例えば、要約書の場合では「本発明」「課題」などの言葉がよく見受けられるが、「分かち書き」する際には不要な言葉である。そのため、「分かち書き」は単語フィルタをかけながら、何度も行い、不要な言語を削除していく。その結果、頻度が高い単語については図４に示す通りである。このうち、「水素」「ガス」「燃料電池」は主語になる名詞、「備える」「有する」「設ける」は動詞、「提供」「供給」「充填」「貯蔵」「接続」は「する」を加えることにより動詞になり得る名詞（以下「サ変可能名詞」という。）である。

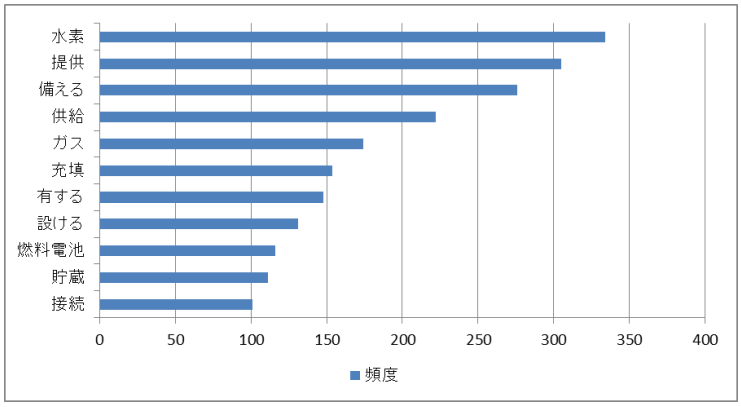

図４：燃料電池ステーション（Ｆターム：5H127FF20）に該当する特許文献のうち要約書で出現する単語の頻度図

３．３．４　「係り受け」と係り受け頻度解析

続いて、係り受け頻度解析を行う。先ず、「係り受け」作業を行う。係り受けは、主語－述語の関係、又は、修飾語－被修飾語の関係をいうが、今回は名詞（サ変可能名詞を含む。）、動詞に絞ってあるため、今回は主語－述語の関係のみをいう。但し、動詞のみならず、前述したサ変可能名詞も述語となり得る。以上より、「係り受け」作業を行い、さらに、頻度が高い主語からどのような述語を受けるかを分析すると以下の結果になる（図５参照）。

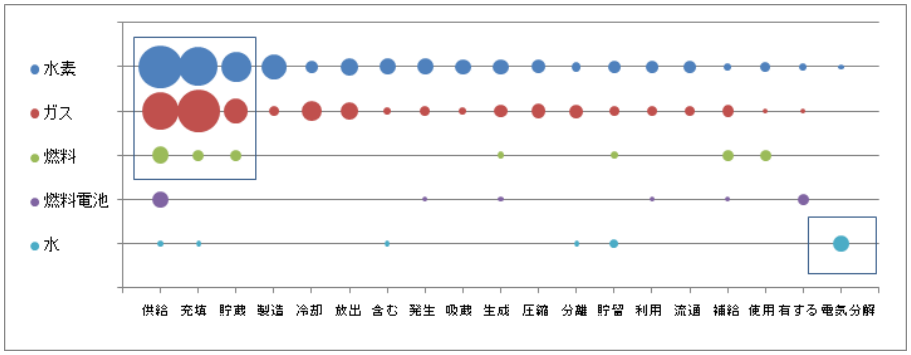

図５：燃料電池ステーション（Ｆターム：5H127FF20）に該当する特許文献のうち主語上位５件の係り受け述語の頻度図

１位の「水素」は２位「ガス」の下位概念であり、２位「ガス」は３位「燃料」の下位概念である。上位下位概念の関係にあるものこれら３語が受ける上位頻出単語は、「供給」「充填」「貯蔵」で一致する。一方、「燃料電池」については「水素」、「ガス」、「燃料」と同様に「供給」は最頻出単語であるが、「充填」「貯蔵」については頻出回数が０である。また、燃料電池の排出物である「水」を受ける単語の最上位頻出単語は、「電気分解」である。このように、係り受け頻度解析を行うことで、データマイニングでは得られない情報を得ることができる。

また、図４の左上側に囲んだ９つの係り受けの年度別頻度推移は以下の結果になる（図６参照。）

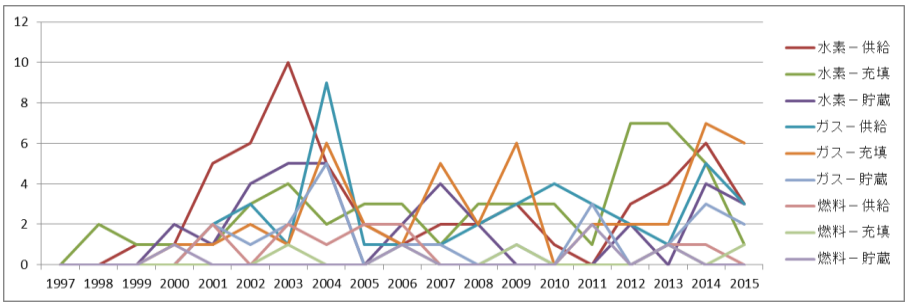

図６：燃料電池ステーション（Ｆターム：5H127FF20）に該当する特許文献のうち係り受け件数上位９個の年度別頻度推移図

３．３．５　話題抽出

続いて、話題抽出を行う。話題抽出は、属性とことば、また ことば同士の関連性の強さをことばネットワーク図で図示します。関連性の指標として、単語同士の「係り受け」関係又は「共起」関係の確率を用いることができます。「係り受け」関係については前述しているので、ここでは説明しないが、「係り受け」関係を指標とした場合のことばネットワーク図を図７に示す。

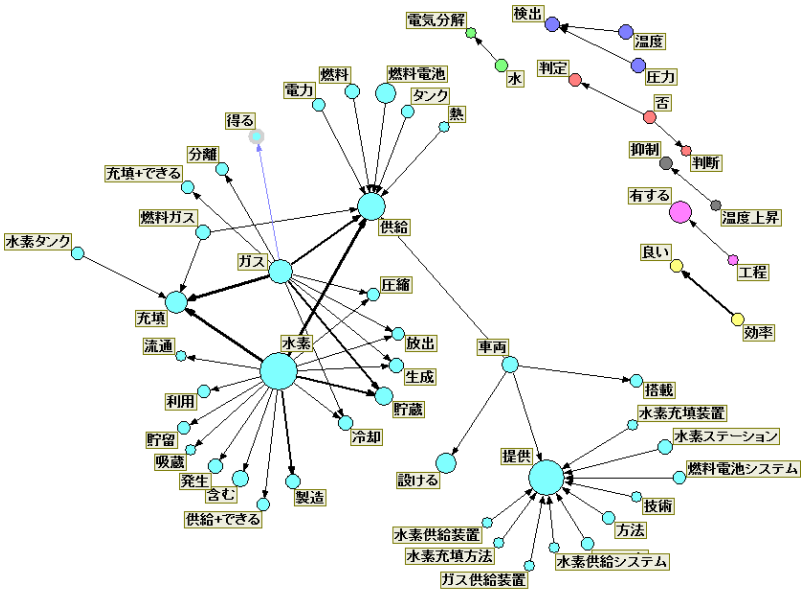

図７：燃料電池ステーション（Ｆターム：5H127FF20）に該当する特許文献の係り受け関係に基づく、ことばネットワーク図

図７のことばネットワーク図は図５の係り受け頻度図と表現方法を変えたもので、図５の係り受け頻度数で上位にヒットした「水素－供給」、「水素－充填」、「水素－貯蔵」、「ガス－供給」、「ガス－充填」、「ガス－貯蔵」の間は太線で示されており、これらの関係が上位頻度であることが分かる。また、このネットワーク図で「水素」、「提供」、「供給」、「ガス」、「充填」の円が他の言葉に比べて大きい。これらの言葉は、図４の単語の頻度図では、それぞれ、１位、２位、４位、５位、６位に該当する。これらの言葉は何らかの言葉と係り受けの関係を有することがこのネットワーク図で分かる。３位の「備える」は単語の頻度数は多いが、ある特定の言語に係り受けをしておらず、多くの言語に係り受けをしていることが考察できる。

「係り受け」関係から、少数意見を拾い上げる。図４で上位５件にヒットした「水素」「提供」「備える」「供給」「ガス」を除く単語の係り受け関係に基づく、ことばネットワーク図を図８に示す。ここでは「安全性」「耐久性」「最小化」などの性能を表す単語が登場する。少数ではあるが、「安全性－向上」「安全性－確保」「耐久性－向上」「エネルギー損失－最小化」「エネルギー効率－向上できる」についても考慮されていることが分かる。

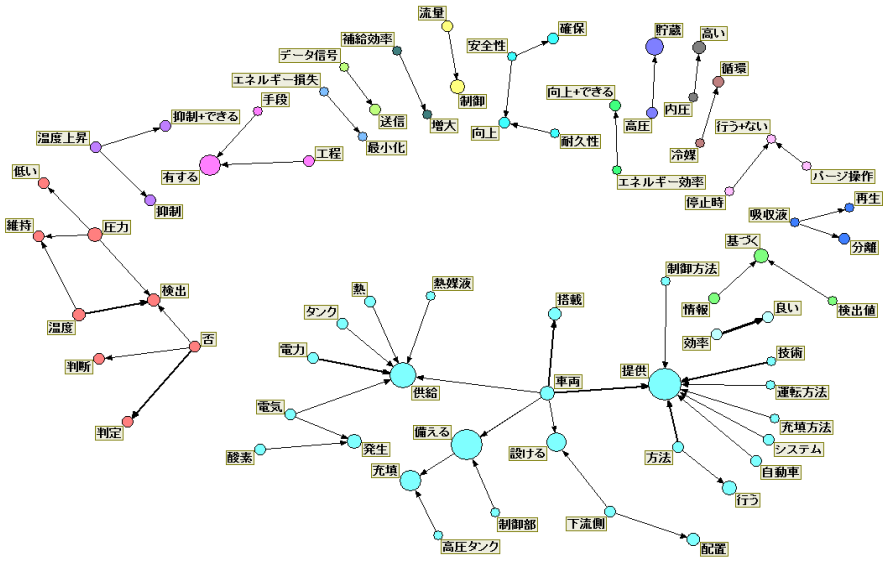

図８：図４で上位５件にヒットした「水素」「提供」「備える」「供給」「ガス」を除く単語の係り受け関係に基づく、ことばネットワーク図

続いて、「共起」関係について説明する。「共起」関係を指標とした場合のことばネットワーク図を図８に示す。なお、「共起」とは、ある単語がある文章（または文）中に出たとき、その文章（文）中に別の限られた単語が頻繁に出現することをいう。例えば、「ワイン」という言葉と「テイスティング」という言葉はよくセットで使われる。また、「ワイン」と「イタリア料理」は高い共起確率を示すが、「ワイン」と「日本料理」はほとんど共起しない（林俊克著「Ｅｘｃｅｌで学ぶテキストマイニング入門」より）。

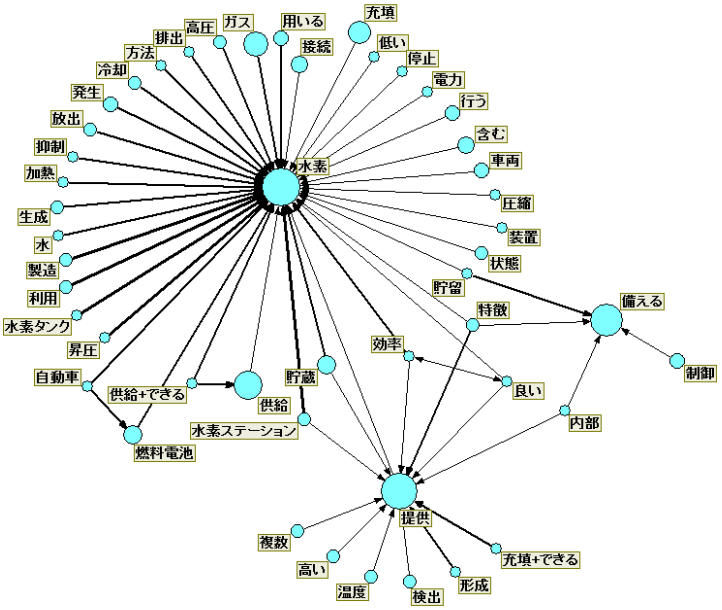

図９:燃料電池ステーション（Ｆターム：5H127FF20）に該当する特許文献の共起関係に基づく、ことばネットワーク図(クラスタ数:１)

「共起」関係の抽出指標を決めるために、単語出現回数を決める。図９図示の例では２０回以上出現したもののみを対象とした。図９のことばネットワーク図は、図４に示す単語頻度数が上位にヒットした「水素」、「提供」、「備える」、「供給」、「ガス」については丸が大きい。「共起」関係は「係り受け」関係とは異なり、修飾－被修飾の関係が無くても良く、同じ文章内にある２つの言葉が出現すれば、それは「共起」関係となる。このことばネットワーク図では、信頼度が高いもの同士が結びつく。

なお、信頼度とは、前提単語が出現したうちの結論単語が出現する確率を言う。例えば、トマト（前提）を買った人のうち、どれくらいの人がきゅうり（結論）も買ったかという確率となる。トマトを買った人は３人で、このうち２人がきゅうりも買っているので、信頼度は（2/3）＊100＝66.7 になる。本件で出現回数が多いのは「提供－水素」で２００件である。但し、前提単語の「提供」の出現回数は３０５であるため、「提供－水素」の信頼度は（200/305）＊100＝65.574 である。一方、信頼度が最も高かったのは「昇圧－水素」である。前提単語の「昇圧」は２６回出現するが、そのうちの２５文章内に「水素」が出現する。なお、「昇圧－水素」の信頼度は（25/26）＊100＝96.15 である。

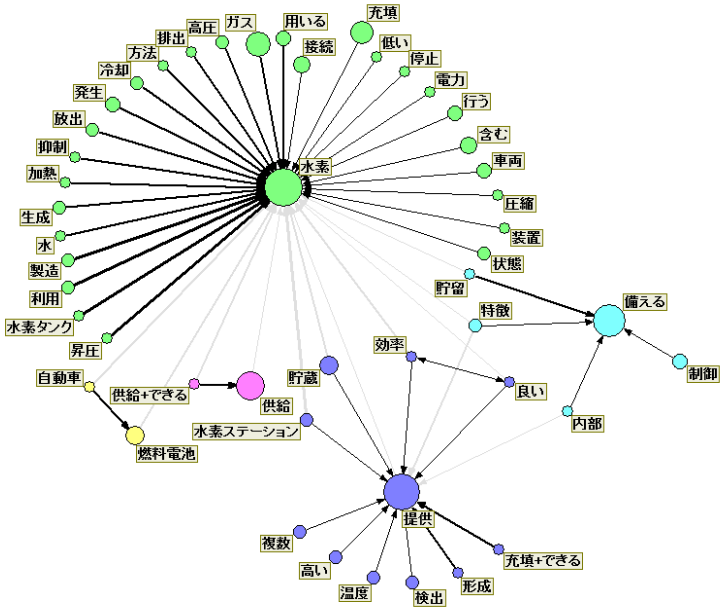

図１０:燃料電池ステーション（Ｆターム：5H127FF20）に該当する特許文献の共起関係に基づく、ことばネットワーク図（クラスタ数：５）

続いて、ことばネットワーク図を整理することにより、関連の強いもの同士の「クラスタ」が見えてくる。なお、テキストマイニングに於ける「クラスタ」とは、話題の集合体をいう。図９ではクラスタ数は１であるが、図１０のように、クラスタ数を５に変えることで、より関連の強いもの同士の「クラスタ」が見えてくる。クラスタ数を５にすることで、話題が「水素」、「提供」、「備える」、「供給」、「燃料電池」であることが分かる。また、単語頻度数で５位の「ガス」は「水素」との結びつきが強く、「共起」関係では話題にならないことが分かる。一方、単語頻度数では９位であった「燃料電池」は「共起」関係では話題になることが分かる。

３．３．６　注目分析

注目分析することで、注目した単語と関連の強い単語を抽出することができる。言い換えると、注目した単語がどのような使われ方をしているか、他にどういった単語と関わりが深いかの把握ができる。今回は単語頻度数では９位であったが、共起のことばネットワークでは５つのクラスタ数の１クラスタになった「燃料電池」を取り上げる。

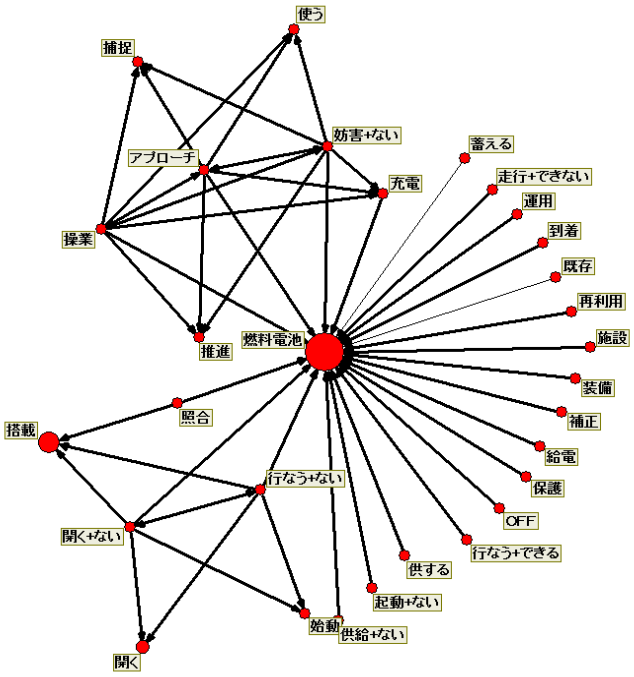

図１１：「燃料電池」に関する注目語情報結果

今回は「燃料電池」を注目語としたが、「アプローチ」、「使う」、「充電」、「推進」、「操業」、「捕捉」、「妨害＋ない」、「操業」がそれぞれ共起しあうことが分かる。同様に、「開く＋ない」、「開く」、「行なう＋ない」、「始動」、「搭載」もそれぞれ共起しあうことも分かる。共起の場合、係り受けよりも広い範囲で単語間の関連を抽出するので、得られた共起語が直接その注目語についてのイメージであるかどうかは不明である。本件においても係り受けのようには考察することはできない。

４．最後に

データマイニングが顧客個人の購買傾向を分析するなどの目的で行われるのに対し、テキストマイニングは顧客個人の特性よりも提供側の状態を把握する面において威力を発揮する。例えば商品の評価や顧客サービスの問題点などを把握することができるという効果がある（「wikipedia」より）と言われている。

今回の事例では、データマイニングから、燃料電池ステーションで取り扱う燃料として、４社とも最も力を入れている「水素」が最も実用的であると考えられる。但し、Ｂ社は「天然ガス等」を、Ｃ社は「ガソリン等」を燃料電池の燃料として採用することも検討していることが図２から判断できると考えられる。

また、テキストマイニングからは、前述したように、図４から判断できることとして、係る単語で同じ概念のもの（例では「水素」「ガス」「燃料」）は、受ける頻度が高い単語（例では「供給」「充填」「貯蔵」）が同じであることがわかる。また。概念が異なる単語（例では「水素」と「水」）は受ける単語も異なる。また、「水素」は「ガス」の下位概念であると述べたが、「水素」は「製造」するものであるのに対して、「ガス」は「補給」するものである。このような違いも係り受け頻度分析でわかる。

以上より、データマイニングとテキストマイニングでは異なった効果を得ることができる。

以上