論　説

キーワードの選定にテキストマイニングを活用した特許検索手法の提案

知的財産情報検索委員会
第　2　小　委　員　会*

抄　録　昨今の技術開発のスピードアップ，技術の深掘りと異種技術との融合による技術の広がり，さらには研究開発と経済活動のグローバル化によって，特に専門外分野，新産業・新市場分野あるいは技術発展が速い分野における高精度の調査が難しくなっていることが分かった。本稿は，検索キーワードの選定プロセスに着目して，従来の特許調査手法にテキストマイニングを取り入れることを提案する。この提案手法は，検索集合の技術俯瞰と，語の関連性把握により，適切な検索キーワードの選定ができることを解説する。

目　次

1．はじめに

2000年以降のパーソナルコンピュータの性能向上，インターネットの普及，WEB技術の開発にともない特許検索データベースのシステム構成が大きく変わり，検索機能が充実するとともに処理能力が飛躍的に向上した。また，各国特許庁と特許検索サービス会社では特許データの電子化が進み，先進国だけではなく新興国のコンテンツがデータベースに収録されるようになり，外国特許調査のニーズ増加に応えるかのように世界各国の特許が手軽に調査できる環境が整ってきた。

さらに，特許検索データベースの進化とともに様々な特許分析ツールが開発されて，特許情報に基づく技術動向調査を活用した出願戦略の立案だけではなく，知財部門と研究開発部門と事業部門が三位一体となった開発戦略や事業戦略の立案にも特許情報の分析が活用されるよう

* 2011年度　The Second Subcommittee, Intellectual Property Information Search Committee

になってきた。

このように特許調査のハード，ソフトの環境が整い，多くの調査者が多種多様な特許情報の調査や分析に取り組む事例が増えている。しかし，個々の調査者の経験値は上昇しているにも関わらず「特許調査が簡単になった」とか,「特許調査の精度が良くなった」という印象は少なく，むしろ調査が難しくなっているような感覚さえある。そこで本稿では特許調査の現状の課題を検証し，次に検索キーワードの選択にテキストマイニングを利用した新しい調査手法を提案して，その効果を解説する。

2．特許調査の課題検証

本章では，技術分野として「自律歩行ロボット」を取り上げ，特許分類を主体とした検索精度に関わる課題を明らかにする。

2．1　技術の変化

日本における自律歩行ロボット（２脚，４脚，６脚）の特許出願を分析した技術動向推移を図１に示す。技術開発の創生期は基本技術である自律歩行の重心制御と学習・推論システムを中心とした出願になっていた。やがて通信，音声認識・合成，画像処理などの異種技術の組合せによる機能を充実させる技術開発や，歩行補助という新しい応用分野の技術開発へと移り変わっている。さらに，最近のPCT出願には「自律的に有機物を探してエネルギーに変換するロボット」[1]という内容もあり20年間という短い期間で，「技術の深掘り」と異種技術との融合による「技術の幅の広がり」によって，自律歩行ロボットの技術分野が発展したことが分かる。

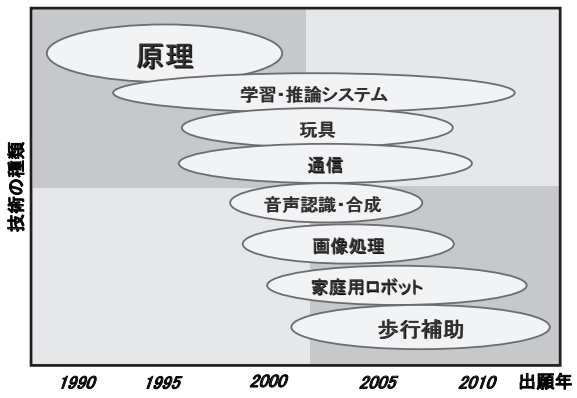

図１　日本の自律歩行ロボットの技術推移

2．2　公報記載IPC[2]

研究開発と経済活動のグローバル化[3),4)]を鑑みて，特許公報に付与されている世界共通の国際特許分類（IPC）の付与精度を確認するために，日本，米国，欧州，中国における自律歩行ロボットの公報記載IPCの付与状況を分析した。図２は縦軸に公報記載IPCのサブクラス，横軸に発行国，バブルサイズを各国におけるそれぞれの特許分類が付与された公報件数としたグラフである。なお，検証に使用したテストコレクション[5]は，各国特許の検索結果をスクリーニングして得られた自作のデータセットである。

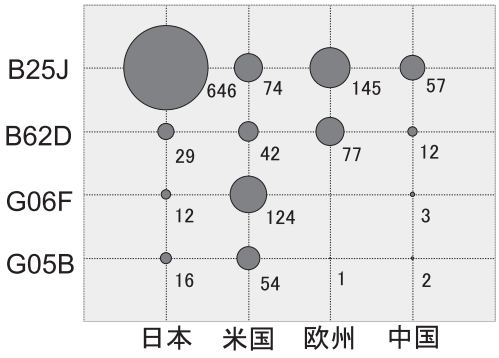

図２　発行国と公報記載IPC（サブクラス）の状況

図２を見てみると，日本では自律歩行ロボットの主要技術分野にあたるB25Jに付与が集中している傾向にあるが，諸外国はB25Jだけではなく，類似技術分野のB62Dや，制御に関する技術分野のG06F，G05Bにもかなりの比率で付与があり，各国の公報記載IPCの付与に大きなズレがあることが分かる。このような各国特

許庁間の付与ズレは，出願内容がほぼ同じであるシンプルファミリ[6]内でも容易に確認できるほど多数の事例を見つける。本稿では一例の分析のみ紹介したが，技術融合が多い機械・電気・通信に関連する分野では多くの事例を見つけることができる。

以上のことから，技術には明確な境界がないため，国際ルールで統一されているIPCであっても，「IPCの解釈」，「IPCの付与基準」，あるいは「発明の訴求ポイントの技術解釈」のいずれかに各国特許庁間の違いがあると公報記載IPCの付与ズレが発生すると考えられる。また，米国では米国特許分類からコンコーダンスによって機械的にIPCへ変換していることも付与ズレの要因と考えられる。さらに，2. 1で述べたような「技術の変化」だけではなく，時の流れとともに技術の解釈や特許分類の解釈にも変化が生じている可能性や，各国の特許法とその運用による影響を受けている可能性もある。

2. 3 ECLA

このような公報記載IPCの付与ズレ問題を解消するために期待されるのは，IPCに比べて技術分類が細分化され，1つの組織が同一水準で特許公報に付与している欧州特許分類（ECLA）である。ECLAは欧州特許庁（EPO）のDocDB（ワールドワイドの書誌検索データベース）上で付与が公開されるが，公報には記載されないデータベース上の利用を前提とした新しい考えの特許分類で，各国特許庁が付与している公報記載IPCより高い付与精度が期待できる。なお，ECLAはPCT，欧州および米国公報を対象に付与され，その他の国（非英語圏）のシンプルファミリに自動的に関連付ける仕組みになっている[7)~9)]。

DocDBにおける各国の公開特許公報のECLA付与率年次推移を図3に示す。全体的に近年にはタイムラグと考えられる付与率の低下が見られる。近年を除くと欧州と米国は95%以上と付与率が高く，その一方で非英語圏の国は欧州・米国に比べると付与率が低く，国によって付与率が大きく異なる。

このような国によるECLA付与率の違いは，PCT出願比率や居住者・非居住者の出願比率に差[10)]があることの影響が大きいと考えられる。そもそも，非英語圏の中でも居住者の内国出願が多い国や，PCT・欧州・米国への出願のシンプルファミリの割合が少ない国はECLAの付与率が低いため，ECLAだけを用いた検索式では致命的な検索漏れが生じることは明らかである。このため，非英語圏の網羅的な調査をするときに安心してECLAが利用できるとは言い難い。

以上はECLAの課題について説明したが，EPOはIPCの更新情報を作成してDocDBに蓄積している。このIPCとECLAは特許分類のコード体系が異なるだけで公報への付与の仕組みが同じため，共通の課題がある。

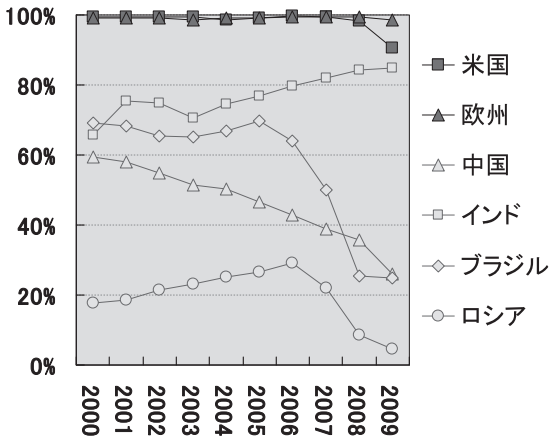

図3　各国のECLAの付与率年次推移

2. 4 特許分類のメンテナンス

特許分類の信頼性を維持するためには，適切なメンテナンスが必要不可欠である。具体的には「新技術に対応する特許分類の更新」（REFORM）

と「バックログ文献への再付与」(UPDATE)が欠かせない。しかし，これらのメンテナンスには相応の手間と時間が必要であり，特に新産業・新市場分野あるいは技術発展が速い分野では調査の精度（再現性・適合性）に影響を与えることが懸念される[11]。

2. 5　課題のまとめ

以上の課題検証により，昨今の技術開発のスピードアップ，異種技術の融合，さらに研究開発と経済活動のグローバル化により，付与のズレ，付与率のばらつき，インデックスの劣化など，特許分類に種々の問題があることが分かった。調査スキルが高いベテランの調査者や，調査者の専門分野であれば，このような問題は苦もなく解決できるかもしれない。しかし，経験が浅い調査者や，経験がある調査者でも技術動向を把握できていない専門外分野，新産業・新市場分野あるいは技術発展が速い分野を調査するときに，検索漏れの少ない高精度の検索結果を得ることが容易いことではないことは明白である。

そこで，技術発展の中で特許調査をする私たちにできる対策は，特許分類を主体とした検索手法に加えて，適切な検索キーワードを利用した解決方法を見出すことであると考え，「検索キーワードの選定」に着目し，特許調査の精度（再現率，適合率）向上を目指すことにした。

特許分類と検索キーワードを併用する検索手法は，従来から知られていることである。しかし，検索キーワードの選定手法については決定的なものが無く，調査者の知見やノウハウによるところが大きい。私たちはキーワード選定のための一手法としてテキストマイニングを利用して，客観性のある機械的なアプローチで検索キーワードの選定にチャレンジした。

なお，本来であれば提案手法の改善効果を推定再現率[12]の数値データを示して検証する必要がある。しかし，人手が多く介在する「検索キーワードの選定」と「検索式」に関して，ヒューマンエラーを取り除いた信憑性が高い検証方法を探すことができなかった。このため，本稿では私たちの検証事例の中から大きな効果が得られた具体例の紹介と解説に留まっていることをお断りして，調査手法の提案に移る。

3．提案する調査手法

利用シーンは，先行技術調査だけではなく，技術動向調査，SDI[13]による調査，他社権利調査など幅広い特許調査を想定している。

3. 1　全体構成

本稿で提案する調査手法（図４）は，従来の調査手法にテキストマイニングによる「技術俯瞰」や「抽出された語の探索」のプロセスを追加した簡単な構成である。作業の流れは，はじめに検索集合の「要約」や「請求項」によるテキストデータをテキストマイニングにより技術俯瞰して，さらに抽出された語の中から適切な検索キーワードを選定して，特許分類と検索キーワードで構成された検索式の作成または見直しをする。

ここで，入力する検索集合の作り方について２つの方法をご紹介する。

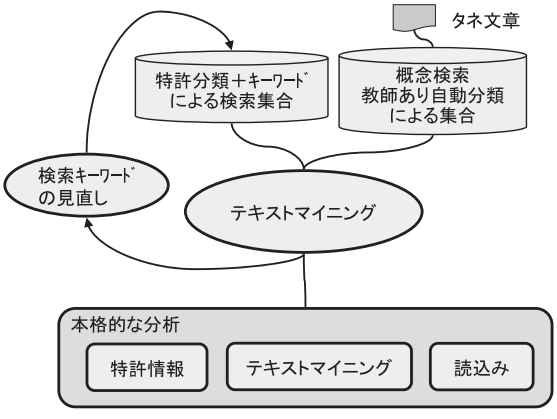

図４　提案手法

（1）従来からある特許調査手法の特許分類と検索キーワードの組合せによる検索集合

（2）「概念検索」[14] あるいは「教師あり自動分類」[15] によって得られた検索集合

概念検索は，注目するニュースリリース，あるいは注目する特許公報などを質問文にした検索結果の類似度上位から得られた検索集合である。一方の「教師あり自動分類」は，既に用意されている特許集合に対して，タネ文章を仕分けの手本とした自動分類により仕分けした検索集合である。

はじめに用意する検索集合は，知見のない技術分野のときはノイズの少ない集合が好ましい。もし知見のある分野であればノイズが多い集合でも問題はなく，ノイズに含まれる他の技術との関連から新たな気づきを得るヒントにすることもできる。

なお，分析対象となる検索集合のテキストデータは，利用する特許検索データベースのダウンロード機能によって得ることができる。調査環境と分析目的によって「要約」を使うか,「請求項」を使うかなど，テキストデータを使い分けると良い。

3. 2　データへのアプローチ

１つの検索集合を複数の観点で分析すると多くの気づきを得られるだけではなく，テキストマイニングの性能を引き出すことに繋がるため,「検索集合全体」,「出願人別集合」,「出願時期集合」の３つの観点で分析すると理想的である。

「検索集合全体」で分析すると技術を俯瞰し易い。しかし，検索集合が大きすぎるとテキストマイニングの可視化処理での「語の表示数制限」により，マップに技術が表現されないことがある。このようなときは，大きな検索集合を小さくすることも兼ねて,「出願人別集合」で分析すると出願人による語の使い方の違いを探索することができる。あるいは,「出願時期集合」で分析すると流行の語を探索することができる。もし，「出願時期集合」の調査対象期間の設定が難しく感じる場合は，最近の特許検索データベースに搭載されている検索結果の簡易統計分析機能や，特許マップ作成ツールを使うなど工夫をすれば容易に適切な調査対象期間を知ることができる。

例えば，自律歩行ロボットにおける出願人別出願件数時系列推移（図５）では，1995年～1999年，2000年～2005年，2006年～2010年の３つの技術開発が盛んな時期があったと考えて，期間Ａ，期間Ｂ，期間Ｃを調査対象期間として分析を進めると良い。

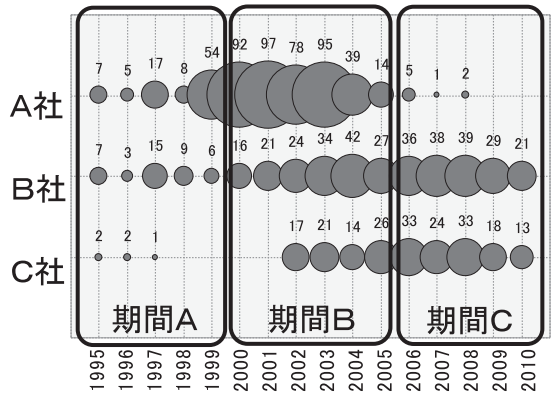

図５　自律歩行ロボットの出願人別の出願件数時系列推移

4．テキストマイニングの詳細

提案手法を多くの方に使っていただけるように，オープンソースのシステムをバックエンドに使ったフリーソフトの「KH Coder」[16] と，オープンソースで統計処理とグラフィック表示が得意なプログラミング言語のフリーソフトの「Ｒ言語」[17] を使って検証を行った。全体の構成を図６に示す。

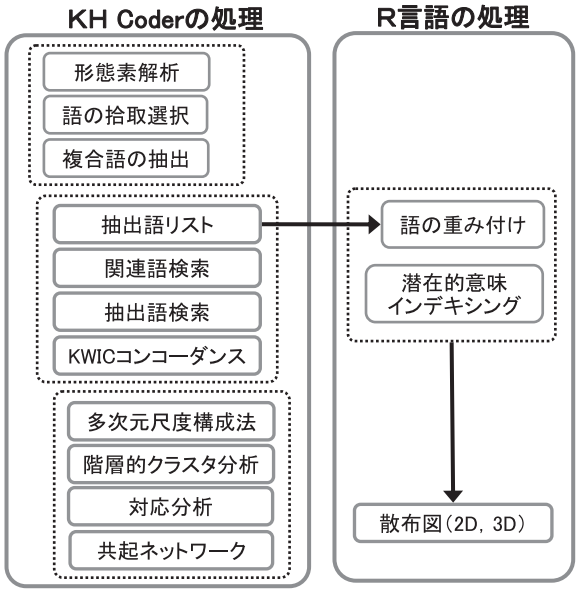

図６　提案手法の全体構成

KH Coderに入力できるテキストデータの言語は日本語と英語であり，機能ブロックは大きく「前処理部」,「語の関連性確認部」,「可視化部」の３つに分けられる。

(1)「前処理部」は日本語と英語に対応した形態素解析[18]，語の取捨選択と複合語抽出の機能を備えている。

(2)「語の関連性確認部」は抽出された語のリスト表示，抽出された語の関連性や文中の使われ方を探る機能として，抽出語リスト化，関連語検索，抽出語検索，KWICコンコーダンスの機能を備えている。

(3)「可視化部」は技術を俯瞰する機能として，多次元尺度構成法[19]，階層的クラスタ分析[20]，対応分析[21]，共起ネットワークの４つの分析アルゴリズムと可視化機能を備えている。

KH Coderでは特許公報の散布図を描くことができないため，５番目の可視化機能としてR言語を使って２次元，３次元の散布図を作成して分析を試みた。

本稿がテキストマイニングに求めるものは，美しい分析マップ作成ではなく，明細書に記載されている語の分析である。このため一般的なテキストマイニングに搭載される同義語辞書や類義語辞書は必要としない。それでは「辞書が無いのに技術把握ができるのか？」という疑問があるだろうが，語が異なるだけで使われ方は同じはずのため共起関係は成立するので大きな問題にはならないと考えられる[22]。

5．分析事例

5. 1　抽出語リスト・複合語の検索

(1) 抽出語リスト

形態素解析の結果を，「抽出語リスト」として品詞ごとに多く出現した順のリスト（図７），または品詞に関係なく出現頻度順に１列に並べたリストを表形式で確認することができる。

名詞	出現数	動詞	出現数	形容詞	出現数
方向	1630	基づく	979	高い	274
関節	1551	行う	855	大きい	177
情報	1130	応じる	462	小さい	104
部材	1071	用いる	368	少ない	94
機構	1045	介す	364	長い	78

図７　品詞ごとの抽出語リスト

また，抽出された語の統計分析機能「記述統計」の「出現回数の分布」を使うと，出現回数が何回以下の語を分析から省けば良いのか？を把握することができる。調査対象のテキストデータから抽出された語と，それぞれの語の出現回数を知ることは，検索や分析を成功させるために重要なことである。

(2) 複合語の検出

例えば，「脚」,「式」,「移動」,「ロボット」のように，形態素解析により語が細かくなってしまった場合に，分析対象に使用したい語であるときは何らかの救済が必要である。もし分析

で使いたい複合語があらかじめ分かっているときは「分析に使用する語の取捨選択」の中で「強制抽出する語の指定」に入れると良い。しかし，複合語が分からないときや，形態素解析により複合語が細かく分解されていて確認が困難なときは，「複合語の検出」（図８）により複合語の存在を容易に確認することができる。

図８　複合語の検出

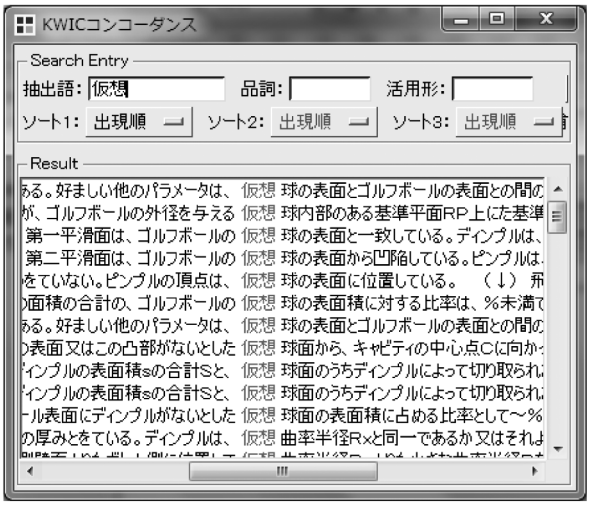

図９　KWICコンコーダンス

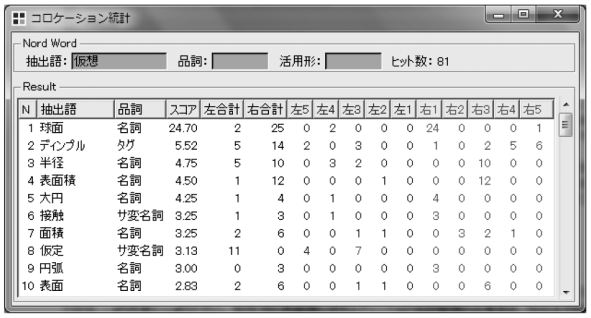

図10　コロケーション統計

5. 2　語の関連性と探索

(1) KWICコンコーダンス

語が用いられている文脈を探るのが「KWICコンコーダンス」（図９）である。検索した語を検索結果の表示画面の中央に配置して，分析対象テキストデータの中で，語が用いられている文脈を探ることができる。また，必要に応じて特定の品詞や，特定の活用形に限定して検索結果をリストに表示できるので，同義語や類義語を探すときにはとても便利である。

さらに，コンコーダンス検索を行った語の周辺に出現している語の集計結果を「コロケーション統計」(図10)により表示することができる。近傍検索の語を探すときや，近傍検索式の距離を決めるときなど，検索式を作るための分析には有効な手段になる。

(2) 関連語探索

注目する語が予め分かっているときは，「関連語探索」を利用すると，その語と強く関連する語を品詞や出現数とともにリストアップし，この結果を使って共起ネットワーク図を描くことができる。

ここで，「関連語探索」の便利な使い方をご紹介する。もし，着目する技術用語の名詞に強く関連する動詞の関係を知りたいときは，「直接入力」に着目する名詞を入力して集計する。その集計結果からフィルタ設定で動詞のみを抽出し，後述する共起ネットワーク図を描かせると名詞と動詞の関係を分析できる。この方法は英文特許を分析するときに有効である。

5. 3　共起ネットワーク図

KH Coderでは，多次元尺度構成法，階層的クラスタ分析，対応分析，共起ネットワークの４つの可視化機能を持っているが，「技術俯瞰」と「語の関連性の探索」を目的としたときに，もっとも相応しいと考えられる共起ネットワークを使って本稿の説明を進める。

共起ネットワークは，語をノードとして共起関係のある語を線で結んだネットワーク図であり，ここではKH Coderの「強い共起関係ほど太い線で描写」と「出現数の多い語ほど大きい円で描写」の２つの機能オプションを選択した例を紹介する[23)]。

(1) 骨の再生医療

再生医療の一つの「微小人工骨ユニットによる骨再生技術」に関する日本特許の要約テキストデータを共起ネットワーク図で示したのが図11である。

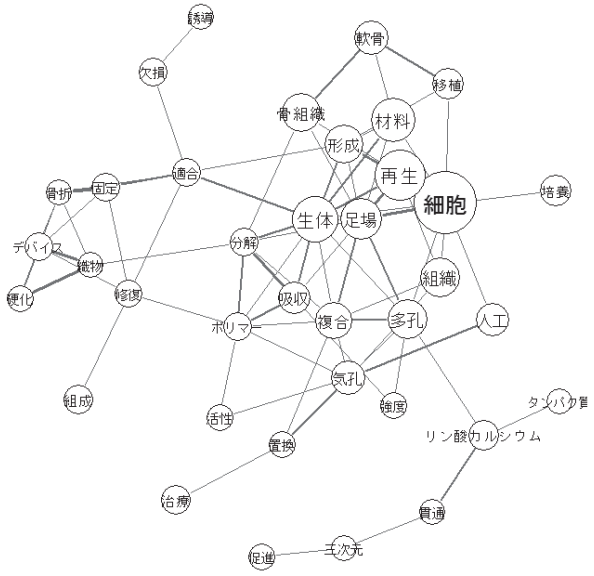

図11　骨再生医療の共起ネットワーク図

この共起ネットワーク図を分析すると，「リン酸カルシウムで作られた組織の侵入性の優れた気孔のある多孔質足場は，分解されて生体に吸収される。」と理解できる。また，この集合の中には骨折治療，軟骨治療などの技術も含まれていることが分かる。

これらの周辺技術は，調査目的から外れるため「ノイズ」と言われることもあるが，技術を俯瞰しながら検索式の見直しをするときに，技術の境界にある大切な情報と考えることができる。また，周辺技術との関連から新たな気づきを得て情報提供ができれば，調査の幅を広げるだけではなく，調査の価値を高めることになる。

(2) 広域ネットワークによる印刷

通信ネットワーク技術の「広域ネットワークによる印刷」に関する米国特許の要約のテキストデータを共起ネットワーク図にしたのが図12である。一般的にソフトウエアや通信ネットワーク技術は，検索集合の作成が難しいだけではなく，苦労して得られた検索集合の分析が難しい技術分野である。

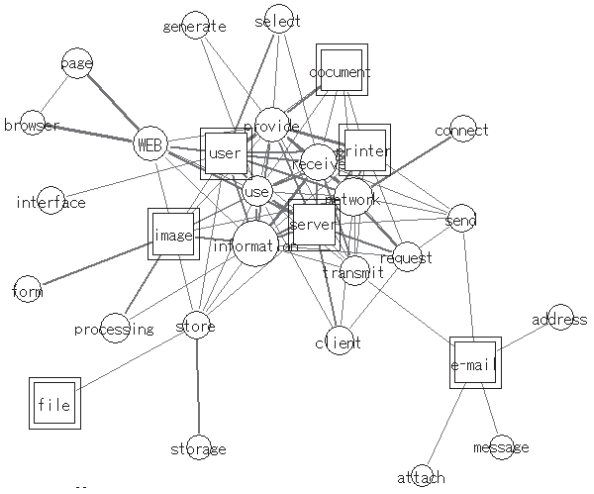

図12　広域ネットワーク印刷の共起ネットワーク図

要約のテキストデータを共起ネットワーク図で分析しただけでは良い結果を得ることができなかった。そこで，着目する語の「document」「printer」「user」「image」「server」「file」「e-mail」

による「関連語探索」の結果を使って共起ネットワーク図を作成した。すると，着目する語が四角の枠で囲まれた通信ネットワークの構造図のような共起ネットワーク図が作成できた。このように技術の骨格を構成する既知の語を手がかりに，語の関連を確認しながら技術を把握することができる。

知見の低い技術分野の調査・分析作業では，はじめに情報を得るのは技術の骨格を構成する語であることが多い。これらの語を「関連語探索」に入力して，語の関連の分析を進めながら技術の理解を深めることができることは，多くの調査者の思考パターンと同じであるため，受け入れやすく効率的に分析できると考えられる。

(3) 自律歩行ロボット

自律歩行ロボットに関する日本特許の要約テキストデータ（1990年～2010年）を共起ネットワーク図で示したのが図13である。既にこの分野に関する知見がある，または技術理解が深まった後に図13を見るとおぼろげに技術を把握できる。しかし，自律歩行ロボットには20年間の技術開発の歴史があり，各時代の技術開発テーマを代表する特徴的な語が表れていないため，検索キーワードになる語を抽出することは難しい。そこで，3. 2で説明した調査対象期間を1995年～1999年，2000年～2005年，2006年～2010年の３つに分割して分析をする。

図14は，1995年～1999年の共起ネットワーク図に，人手で分析した結果の「丸い囲み線」と，「吹き出しのコメント」を付加したものである。この時代には機械と機械の制御に関する技術への出願がある。2000年～2005年の分析結果（図15）では，機械，音声合成・画像認識，経路探索，重心制御に関する出願がある。2006年～2010年の分析結果（図16）では，歩行補助具，移動経路，各種の制御に関する出願があることが分かる。このように調査対象期間により検索集合を分割して，分析することで特徴的な語を浮かび上がらせて，検索キーワードを選定することができる。

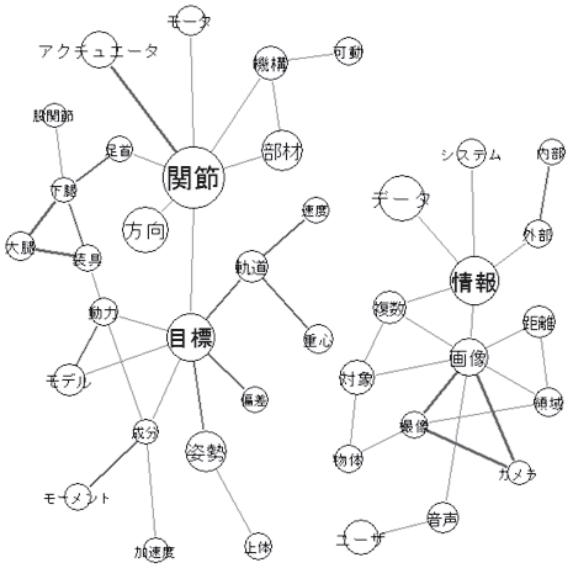

図13　自律歩行ロボット1990年～2010年

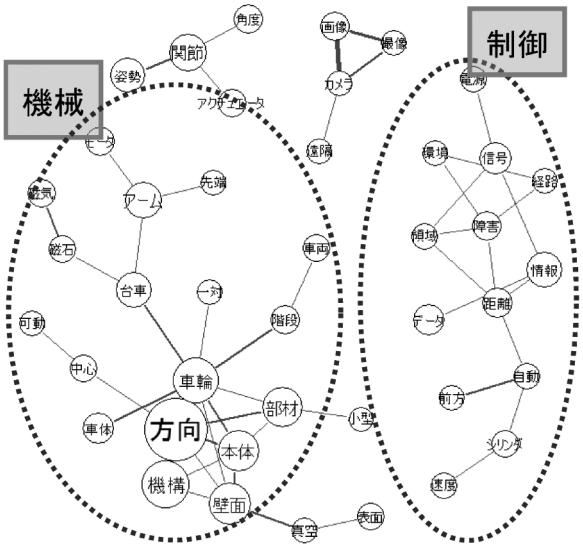

図14　自律歩行ロボット1995～1999年

共起ネットワーク図は，経営層へのプレゼンテーションに向くような綺麗な特許マップではないが，語の繋がりのシンプルなマップにより「マップの解釈の差が少ない」，「信憑性が高い」という特徴を有している。

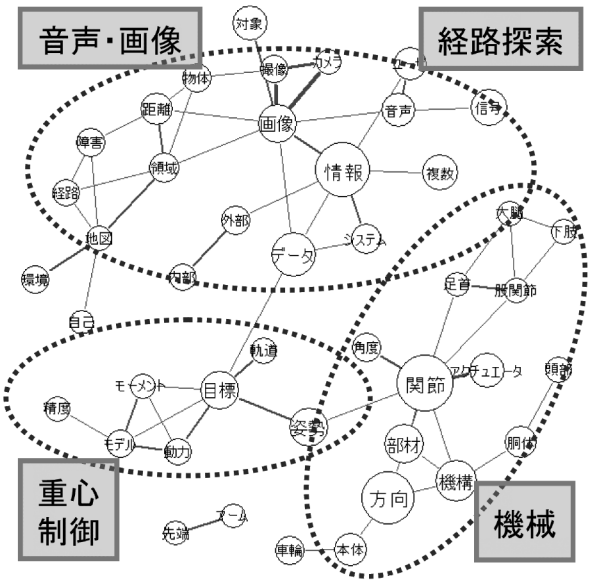

図15　自律歩行ロボット2000～2005年

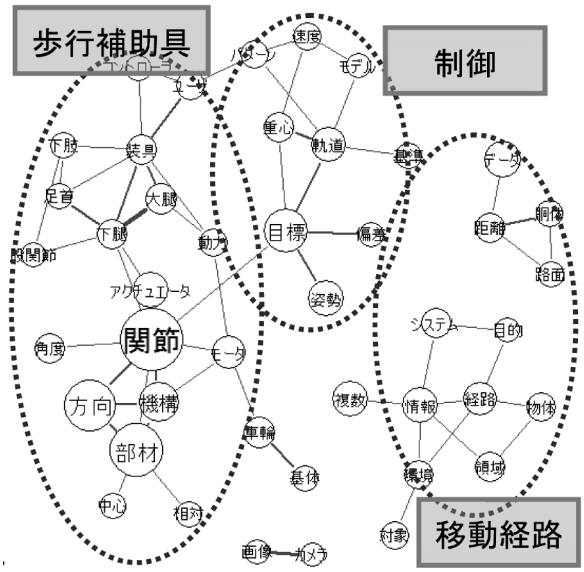

図16　自律歩行ロボット2006～2010年

5. 4　散 布 図

検索集合の技術を俯瞰する方法として，２次元の特許の散布図と，特徴的な特許を抽出するために２次元の散布図のＺ軸方向に情報を加えた３次元の散布図を作成した。

散布図を作成するアルゴリズムは潜在的意味インデキシングを利用した。潜在的意味インデキシングとは文書・抽出語行列を高次元行列から低次元行列に次元圧縮を行って，文書どうしの関連性を明らかにすることができる技術である[24)～27)]。

散布図を作成する作業は，KH Coderの形態素解析などの前処理を済ましたのち，「『文書・抽出語』表の出力」によって文書・抽出語行列を出力する。つぎに，Ｒ言語を使って，文書・抽出語行列をTF・IDF[28)]による重みづけ処理と，特異値分解により得られた左特異行列を用いて文書・抽出語行列の低次元化を行ったのちに，散布図の可視化を行う。図17は上述したＲ言語のプログラムの例である。

```
setwd("C:/R/sample")
library(lsa)
matrix <-read.table(" walking_robot_BIG3.txt ")
matrix_a <- t ( matrix )
weighted<-lw_logtf(matrix_a) * gw_idf(matrix_a)
weighted.svd <- svd(weighted)
result<-t(weighted.svd$u [,1:2]) %*% weighted
result<-t(result)
plot(result)
```

図17　Ｒ言語による散布図可視化プログラムの例

自律歩行ロボットの上位３社の要約のテキストデータを，KH CoderとＲ言語による散布図可視化プログラム（図17）を使って作成した散布図に，人手で分析した結果の「丸い囲み線」と，「吹き出しのコメント」を付加したものである（図18）。Ａ社，Ｂ社，Ｃ社のそれぞれの色を変えて表示させると，Ａ社がエンターテイメント，Ｂ社が基本構成と安全制御と歩行補助，Ｃ社が基本構造や歩行補助という各社の出願傾向があることが分かる。

散布図の点で示される１つの特許の位置を変更しないように，年代によって特許の点の大きさを変えると，時系列推移（タイムスライス）

を描くことができる（図19）。もし，図18のような分析が終わっていれば，この時系列推移と組み合わせると技術の遷移を容易に把握することができる。

また，２次元散布図の中の点で示される１つの特許に「重み」を付けると３次元散布図を描写することができる。Ｚ軸方向に被引用文献数の情報を加えて，さらに，審査状況に応じて１つの点の大きさを変えた例を図20に示す。

「重み」の情報は，被引用文献数，審査状況，引用文献数，パテントファミリ数，請求項数，出願人名称などを数値化して，「Ｚ軸」，「点の大きさ」，あるいは「色」に変換して可視化する。このように，Ｒ言語の多彩な表現力を活用して特許情報を可視化することで，新たな観点で特徴のある特許を探すことができる。

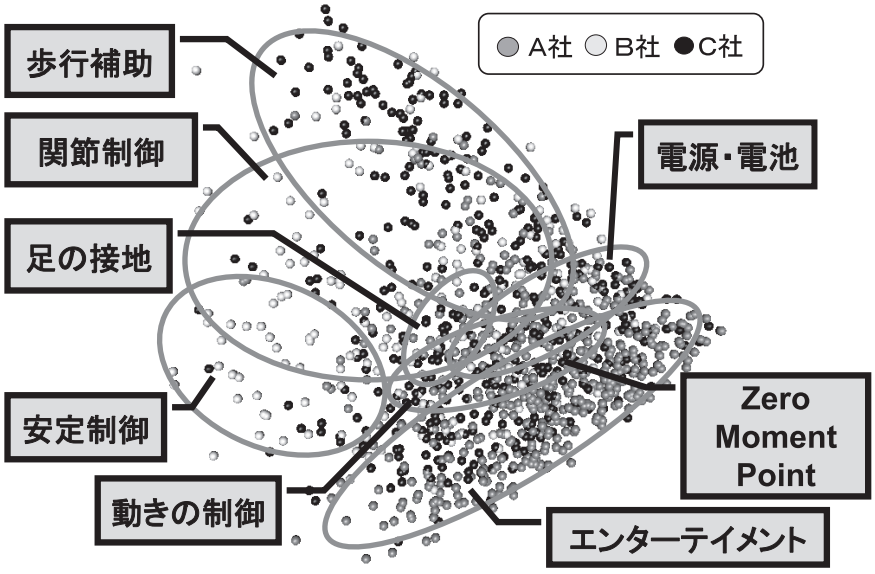

図18　散布図と分析結果

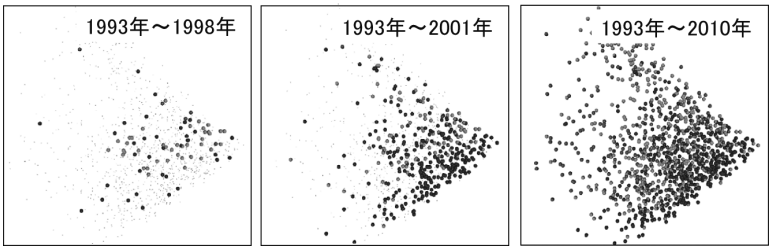

図19　散布図の時系列推移

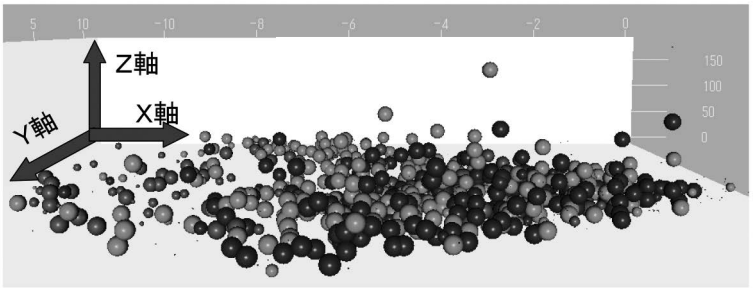

図20　三次元の散布図

以上のように２次元，３次元の散布図を活用して，新しい観点の分析や価値評価に使えるのではないかと検証を行ってきた。しかし，図18においてそれぞれの技術を示している囲み線と，吹き出しコメントの技術名称は，調査者が一つ一つの特許を見て分析したものである。例えば，図18の吹き出しコメントの「エンターテイメント」は発明の訴求ポイントが音声合成，音声認識，音楽などに関わる技術の集まりで，これらの技術を総称した上位概念の名称である。２次元の散布図における技術把握が調査者の知見に依存するだけではなく，調査者による作業が介在するため，調査者により分析結果が異なってしまい信憑性と客観性に乏しいという大きな課題が残ってしまった。

散布図と特徴的な語を組み合わせた可視化技術とユーザーインタフェースのさらなる開発により，今後，信憑性が高い分析ツールになる可能性があるのでこれからの研究に期待したい。

5. 5　分析のコツ

最後に，テキストマイニングによる特許分析で，気をつけたい「分析のコツ」を４つご紹介する。

(1) 出現頻度の低い語の扱い

文書中に出現する語には「Zipfの法則」という経験則があることが知られている。これは『文書中に出現している語をその出現頻度の大きい方から順に並べると，Ｎ番目の語の頻度は一番頻度の大きい語の１／Ｎになっている。』というものである。言語の違いや記述内容の違いにより必ずこの法則に当てはまるとは限らないが，文書中に出現する語は，出現頻度が特に高い少数の語と，出現頻度の低い多数の語から構成されるという傾向が確実に見られる。このため，技術把握をする上では，出現頻度が低い語は分析対象としてあまり意味がない[29]。

(2) 出現頻度の高い語の扱い

文書中に出現頻度が特に高い少数の語が，不要な語である可能性が高いことも知られている。例えば，自律歩行ロボットでは「ロボット」や，ゴルフボールでは「ゴルフ」「ボール」は，出現頻度が極めて高く，対象となる技術分野を直接的に示す語であることから，不要な語として扱うと分析が容易になる。

(3) 可視化による技術把握が上手くいかないとき

技術が成熟していて，特許出願激戦区の「ゴルフボール」のような技術分野では，要約のテキストデータの可視化マップによる技術把握が難しいことがあった。このようなときは技術分野の絞り込みを行う，請求項のテキストデータを分析に利用する，などの対応をすると良い結果を得ることができた。

それでも技術把握ができないときは，「KWICコンコーダンス」または「関連語探索」に戻って，技術の中心となる技術用語を手掛かりに分析を進めると良い。

(4) 何をやってもうまくいかないとき

どんなに優れたツールでも入力したテキストデータの品質が悪ければ分析することはできない。例えば，古いタイプの機械翻訳，化学式の羅列，漢字と外来語のカタカナの同義語・異表記が多数混在するなど，「人間が読んでも理解できないテキストデータ」のときは分析が困難なことが多い。何をやってもうまくいかないときは，前処理の形態素解析の結果である抽出語リストにより，入力したテキストデータの品質を確認すると良い。

6. ま と め

技術開発を取り巻く環境の変化によって，専門外分野，新産業・新市場分野あるいは技術発

展が速い分野を調査するときに，特許分類主体の検索だけでは高精度の調査を行う際に課題があることが分かった。このため検索キーワードの選定に着目して，従来の特許調査手法にテキストマイニングを取り入れることを提案した。この提案手法を使うと「抽出された語と複合語のリスト確認」，「語の関連性や語の使われ方の探索」，「語の関連性の可視化」により技術を俯瞰して，適切な検索キーワードを調査者の知見やノウハウに頼ることなく選定し，検索式に加えることができるため，専門外分野，新産業・新市場分野あるいは技術発展が速い分野を調査するときには，調査の高精度（適合率，再現率）化に有効であると考えられる。ただし，今回の検討は5．5（4）に記載したような課題があることも事実であり，あらゆる技術分野，あらゆる品質のテキストデータについて一元的に取り扱うことができるか否かまでは詳細な検証はできなかった。今後，本稿で扱った事例以外の検証が進むことを期待したい。

7．おわりに

2010年の世界の特許出願件数は197万件になった[30]。新興国の経済発展と，研究開発と経済活動のグローバル化が進み，特許出願件数はこれからも毎年増え続けると予想され，世界中の国々で人手による適切な特許分類を付与することは益々困難になり，今まで以上に特許調査が難しくなると考えられる。

このような状況の中では，本稿で紹介したようなテキストマイニングを取り入れた調査手法の重要性が増してくると考えられる。しかし，「言語」と「記載内容の品質」はテキスト検索・テキスト分析の永遠の課題として残るため，一つの提案をさせていただきたい。

それは『すべての特許公報の要約テキストデータを世界共通の様式で記載する』ことである。具体的には，要約の言語を「英語」にして，記載事項を「新規性」，「用途」，「優位性」，「実施例」などを基本様式とする。そうすれば英語で世界の特許をテキスト検索できる。さらに，記載事項が統一されることにより記載内容の品質向上が期待できるため，各種特許調査や特許分析の手助けになるであろう。このような抄録は商用データベースにはあるものの，すべての人が利用可能という訳ではない。できるならば公的な機関のもとで無料公開できる「世界共通様式の要約テキストデータ」の実現を希望する。

本稿が読者のみなさまの特許調査手法の見直しのヒントと，知財情報をもとにした技術動向分析による知財マネジメントの一助になれば幸いである。

なお本稿は，2011年度知的財産情報検索委員会第２小委員会第１WGメンバーである谷口誠一（小委員長　セイコーエプソン），荒瀬真理子（ダンロップスポーツ），川角容子（リコー），鷺谷喜春（富士ゼロックス）および武島正治（積水化学工業）の執筆による。

注　記

1） WO2010/053957A1

2） 特許公報に記載されたIPCのこと。一方，特許公報に記載されないデータベース上の特許分類には，欧州特許庁の更新IPC，ECLA（欧州特許分類），日本特許庁の整理標準化データによる更新FI，更新Ｆタームなどがあり，サービスを提供しているデータベース会社が独自に収録している。

3） WIPO「World Intellectual Property Indicators, 2011 edition」
Patent applications by patent office and filing route, broken down by resident and non-resident（1995-2010）
http://www.wipo.int/ipstats/en/statistics/patents/（参照日：2012年４月24日）

4） Statistics on World Population, GDP and Per

Capita GDP, 1-2008 AD（Horizontal file, copyright Angus Maddison, university of Groningen）
http://www.ggdc.net/MADDISON/oriindex.htm（参照日：2012年４月24日）

5） 情報検索システムの性能評価に用いる正解データを含めた実験用データセットのこと。本稿では自律歩行ロボットの各国の検索集合を人手によるスクリーニングで得られた精度の高いデータセットを用いた。

6） シンプルファミリとは，同じ優先権番号をもつ特許群（パテントファミリ）のことで，欧州特許庁が作成している。Espacenetで表示されるファミリ情報である。

7） 情報管理　2010年８月号「欧州特許分類の理論と活用　国際調和に向かって世界をリードする検索ツール」武藤 晃，村野 祐子，鈴木 智香
http://www.jaici.or.jp/stn/pdf/ref-class.pdf
（参照日：2012年４月24日）

8） STNEWS，Vol.27，No.5，2011年，９・10月号，日本版　編集・発行：化学情報協会
http://www.epo.org/searching/essentialls/patent-families/espacenet.html
（参照日：2012年４月24日）

9） The Espacenet patent family（欧州特許庁）
http://www.epo.org/searching/essentials/patent-families/espacenet.html
（参照日：2012年４月24日）

10） 文部科学省：「科学技術・イノベーション政策の展開にあたっての課題等に関する懇談会」これまでの議論の取りまとめ
http://www.mext.go.jp/b_menu/shingi/chousa/gijyutu/014/houkoku/1283136.htm
（参照日：2012年４月24日）

11） 特許庁検索システム最適化調査報告書（日本特許庁）
5．自動分類付与技術に関する調査
5－2．検証項目
5－2－1．想定する業務適用場面
（1）FI改正・Ｆタームメンテナンスの必要性と課題
http://www.jpo.go.jp/shiryou/toushin/chousa/pdf/kensaku_saitekika/kensaku1_5.pdf
（参照日：2012年４月24日）

12） 推定再現率とは，検索結果の網羅性のことで，全特許情報の中で，検索目的の正解特許件数を（Ｒ）を分母として，自分が見つけた特許件数（Ｐ）の割合（100＊Ｐ／Ｒ）である。
参考文献「特許分析・解析の哲学小道」桐山勉 情報管理，Vol.52，No.5

13） SDIとは，Selected Dissemination Informationの略。
ある目的の調査をするために検索式を作成して，定期的に特許調査を行って情報収集すること。

14） 概念検索とは，任意の文章を入力して，その文章の内容に類似する文書を類似する度合い（類似度）の高い順に出力する検索方式のこと。

15） 教師あり自動分類には大きく２種類あり，一つはＦ値という統計的手法により類似性の高い特許を仕分ける方法。もう一つは，各文書の内容を文書中に出現する単語の頻度に基づいて単語ベクトルと呼ぶ形式で表現し，文書間の類似度を，ベクトルの余弦などで定義して仕分ける方法がある。
参考文献①「テキストマイニング技術の特許データへの適用」林田英雄
UNISYS TECHNOLOGY REVIEW 第87号 NOV，2005
参考文献②「企業の情報と知識の利活用を促進する対話型文書分類システム」
東芝レビュー　2010，Vol.65，No.2

16） KH Coderのホームページ
http://khc.sourceforge.net/

17） Ｒ言語のホームページ
http://www.r-project.org/

18） 文書データは基本的に連続した文字列で構成されている。形態素解析は，この一連の文字列を文法的に意味のある単位の構成要素に分割し，各要素の文法的素性を決定すること。
参考文献「テキストマイニングの自然言語処理」那須川哲哉　21ページ

19） 多次元尺度構成法とは，各個体間の類似の度合いや非類似の度合いを，それぞれの個体から測定した多変数の数値から求め，その結果を２次元または３次元に描写して，固体間の関係を探る手法のこと。
参考文献「テキストマイニング入門」石田基広 151ページ

20） 階層的クラスタリングとは，基本的には距離行列を用いて，似ているものを段階的にグルーピ

ングする。樹形図を用いてマップを作成する手法のこと。
参考文献「ESTRELA」2009年３月（No.180）

21） 対応分析とは，クロス表を対象に行と列を調整して，相関が強いもの同士が隣接するように並び替えを行う分析手法のこと。
参考文献　石田基広，Ｒによるテキストマイニング入門，163ページ，2008，森北出版

22） 那須川哲哉，テキストマイニングを使う技術／作る技術，172ページ，2006，東京電機大学出版局

23） ３つの事例紹介で示した共起ネットワーク図は本稿スペースと説明のし易さを優先して，表示する語を少なく表示しているが，表示設定の変更により多くの語の表示が可能である。

24） 那須川哲哉，テキストマイニングを使う技術／作る技術，172～175ページ，2006，東京電機大学出版局

25） 石田基広，Ｒによるテキストマイニング入門，136～137ページ，2008，森北出版

26） 北研二，津田和彦，獅々堀正幹，情報検索アルゴリズム，65ページ，2002，共立出版

27） 「ESTRELA」2009年７月（No.184）

28） TF・IDFとは，ある単語が文書中でどの程度重要かを識別するための指標のこと。
tf（term frequency）は，その文書内の単語の重要度のことで，その文書内で多く出現するものを重要とみなす。
idf（inverse document frequency）は，他の文書を考慮したときその単語がどの程度特徴的（≒重要）かを表し，多くの文書に出現するものは重要でないとみなす。

29） Christopher D. Manning, Prabhakar Raghavan, Hinrich Schuetze,「Introduction to Information Retrieval」, 369～384ページ, 2008, Cambridge University Press

30） WIPO 2010 World Intellectual Property Indicators
http://www.wipo.int/ipstats/en/wipi/index.html（参照日：2012年４月24日）

（原稿受領日　2012年７月10日）