富士通知財ソリューション「ATMS」の検索、分析技術

（株）富士通研究所　ソフトウェア＆ソリューション研究所　　渡部　勇
（株）富士通長野システムエンジニアリング　特許ソリューション部　待井　学

1. はじめに

1.1. 富士通の知的財産ソリューション

1.1.1 知的財産部門の課題

「知財が経営を左右する」と言われ、より知的財産の重要性が増す中、企業の知的財産部門は今後何を強化すべきと考えているのだろうか。2007年に当社で実施した知財戦略セミナーで約200名にアンケートしたところ、表1のような結果を得ることができた。

大別すると、将来事業への貢献（いわゆる攻めの部分）と発明発掘や侵害回避（いわゆる守りの部分）の両面に重点を置いていることがわかる。

1.1.2 攻めと守りの知財システム

それでは、こういった攻めの部分と守りの部分をバランスよく強化していく理想の知財システムとはどんなものなのか考察してみる。

守りの部分でいうと、必要な人が、すぐに使える環境において、簡単な操作、すばやいレスポンスで高精度なデータを入手できることが重要である。つまり、特許管理・調査業務を最大限効率化するシステム構築が必要である。このステージを「基盤構築ステージ」と呼ぶ。

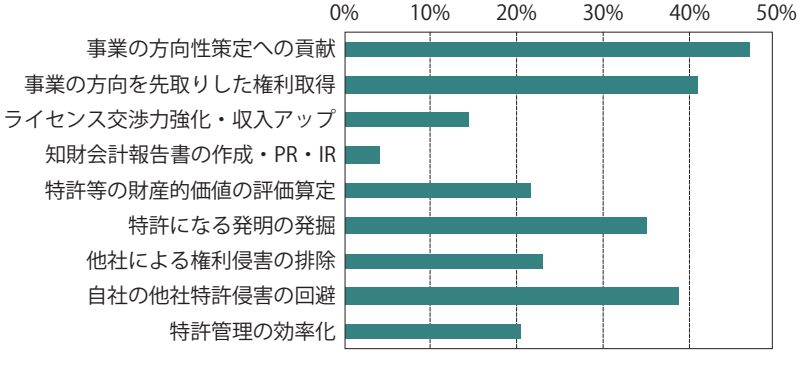
一方攻めの部分でいうと、知財の観点から、事業や研究開発部門に対して、戦略策定の判断材料となる高精度なデータをすばやく提供できることが重要である。つまり、特許分析・可視化するシステム構築が必要である。このステージを「情報活用ステージ」と呼ぶ。

現状、基盤構築ステージを構成する特許管理システム、特許検索システムを構築済の企業は殆どであるが、情報活用ステージを構成する特許分析システムなどを構築している企業はまだ少数である。

1.1.3「ATMS」知的財産ソリューション

富士通では1980年代より、社内外の知財を管理、調査するシステム「ATMS」（アトムズ）を販売している。これは富士通社内で利用してきたシステムを外販したものである。

表1　企業が知財に対して今後強化すべきと考える点（富士通知財戦略セミナーアンケートより）



現在では、特許管理システム（ATMS/PM2000）、特許検索サービス（ATMS/IR.net）、特許出願支援ソフト（ATMS/PPW）に、特許分析システム（ATMS/Analyzer）を加え、知財に必要な業務パッケージをATMSという1つのブランドで提供している（図1）。

富士通では、これらの業務パッケージを組合せて基盤構築ステージ、情報活用ステージを実現する構築ステップを図2で提案する。この構築ステップでは更に戦略展開ステージと呼ぶ、特許情報＋非特許情報から、意思決定支援システムを構築するという将来コンセプトも含む。

今後重要になるであろう情報活用ステージや戦略展開ステージにおいて、なくてはならないITツールのひとつとして、テキストマイニング技術を採用した特許

図1　ATMSソリューションマップ

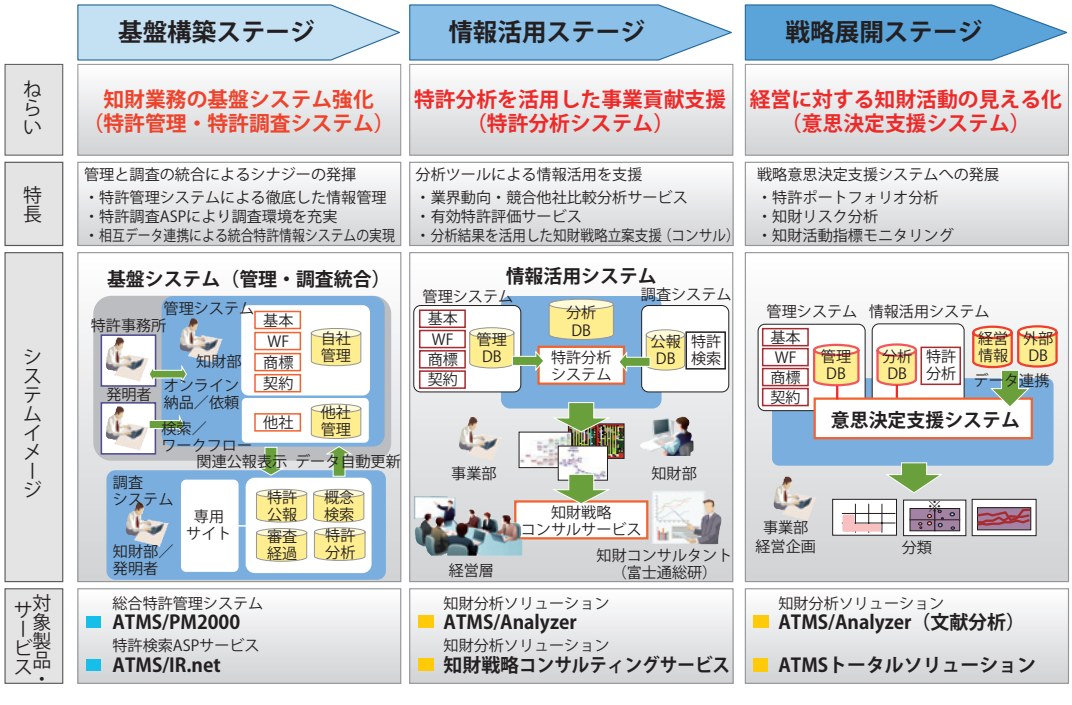


図2　富士通のご提案するシステム構築ステップ

分析ツールがあげられる。富士通研究所ではいち早く、テキストマイニングの技術に取り組み、実用化している。次章以降では、特許検索サービスATMS/IR.netや特許分析システムATMS/Analyzerのベースとなっている当社研究所の最新技術をご紹介していく。

2. テキストマイニング技術

2.1. テキストマイニング技術の概要

テキストマイニングとは、文書情報から有益な知識を発見・抽出するための技術である。情報検索システムが、利用者の目的に合った文書を探し出すことを目的としているのに対し、テキストマイニングでは、文書を個別に調べても分からない、文書群全体に内在する知識（パターンやトレンド）を発見することを目的としている。まだ比較的新しい研究領域ではあるが、この十数年の間に実用化も急速に進み、大量のテキスト情報にアクセスするための新しい道具として、ビジネスの場面でも活用されるようになってきている[1]。

テキストマイニングは、さまざまな要素技術を組み合わせた複合的な技術である。これらの要素技術は、テキストマイニングを進めていく上での情報処理の流れに合わせて、以下の3つに分類することができる。

- 概念抽出技術：自然言語で書かれた文書情報からその内容をあらわす概念を抽出
- マイニング技術：抽出された概念を統計的に分析
- 可視化技術：マイニング結果を人間が理解しやすい形に可視化（視覚化）し、対話的な分析を実現

テキストマイニングシステムにはいろいろなタイプのものがあるが、いずれも上記の3つの要素技術で構成されるという全体の枠組みは変わらない。以下では、テキストマイニングの基本となる概念抽出技術について解説する。

2.2. 概念抽出技術

テキストマイニングを行なうためには、まず分析対象である文書情報から、その内容をあらわす概念を抽出する必要がある。

例えば、フリーアンサー（自由記述式）のアンケート結果を分析する場合、選択式の回答項目に関しては、選択肢ごとに件数を集計してやれば、どのような意見が多かったのかをすぐに調べることができる。一方、フリーアンサーの部分に関しては、同様の集計を行なっても期待するような結果は得られない。自然言語では同一の内容をさまざまな表現であらわすことが可能なため、文字列レベルでの集計を行なっても意味がないからである。文書情報を分析するためには、文書全体の文字列をそのまま使用するのではなく、その内容をあらわす概念を抽出（コード化）し、同一あるいは類似の内容をまとめて集計・分析できるようにしてやる必要がある。

2.2.1 重要単語の抽出

文書情報の内容をあらわす概念を抽出する方法としてまず挙げられるのは、自然言語処理技術を利用して，文書中の重要単語を抽出する方法である。テキストマイニングはもちろんのこと、情報検索・文書分類などテキスト情報を扱うさまざまな分野で利用されている最も基本的なモデルである。

分析対象となる文書情報は、まず形態素解析により単語単位に分割される。この単語群に対し、

- 辞書などを用いて表記の揺れ・同義語を統一
- 品詞情報・統計情報を用いて複合語を抽出[2]
- 特定の品詞[3]の単語を選択
- キーワードにはならない「こと」「とき」などの一般語（不要語）を削除
- 統計量などによって単語の重みを計算[4]

を行なうことで、各文書に対する重要単語（キーワード）

1）富士通研究所では、特許分析のほかに、マーケティング（自由記述のアンケート分析、コールセンターのログ分析、ブログを用いた評判分析）や、リスクマイニング（トラブル情報・障害情報を分析）への適用研究・実用化を進めている。

2）形態素解析では、単語よりさらに細かい形態素という単位に分割される。形態素のレベルでは分析の単位としては細かすぎるため、形態素を組み合わせた単語・複合語レベルの情報を抽出する必要がある。

3）名詞・未登録語などを利用するケースが一般的であるが、分析の目的によっては形容詞などを用いることもある。

4）重み付けとしては、文書中での単語の頻度TF（Term Frequency）と、単語が出現する文書数の逆数IDF（Inverted Document Frequency）を用いたTF・IDFと呼ばれる方式がよく用いられる。他に、相対エントロピー（Kullback-Leibler距離とも呼ばれる）などが用いられることもある。いずれも「文書中に多くあらわれる単語」の重みを大きくし、「多くの文書にあらわれる単語」の重みを小さくするような指標になっている。

のリストが得られる。

この重み付けされた重要単語のリストは、集計・分析の基本単位として利用される。例えば、先のアンケート分析の例で言えば、回答全体における単語の重みを集計することにより、どのような話題・テーマに関する意見が多かったのかを知ることができる。また、概念検索・クラスタリングなどに用いられる文書間の関連度・類似度の計算にも利用される[5]。

2.2.2 係り受け組の抽出

文書の内容を重み付きの単語リストとして表現するというモデルにより、文書の扱う話題・テーマを捉えることはできるが、事実・意見といったより深いレベルの内容を扱うことはできない。

例えば、以下のような3つの文を分析する場合,

・「AはBであり、CはDである」
・「AはBであり、CはDではない」
・「AはDであり、CはBである」

これらの文を単語リストで表現すると、いずれも（A, B, C, D）となり、AやBやCやDに関する話題・テーマを扱っているという表層的な内容を捉えることはできるが、3つの文の違いを捉えることはできない。

上記3つの文の違いを捉えるためには、形態素解析結果に対してさらに構文解析を適用し、以下のように、語と語（あるいは文節と文節）の間の係り受け関係を抽出する。

・「AはBである」「CはDである」
・「AはBである」「CはDではない」
・「AはDである」「CはBである」

上記のように、文書の内容を係り受け組によって表現することにより、単語レベルでは捉えることのできない文意の違いを捉えることが可能となる。

2.2.3 情報抽出

形態素解析や構文解析といった自然言語処理技術に加え、辞書やルールによる情報抽出技術を利用することによって、より深いレベルの意味・内容を抽出する以下のようなアプローチもある。

・係り受け解析とルールベースの情報抽出を組み合わせることにより、障害情報から、障害の現象・原因・対策をあらわす情報を抽出[6]
・係り受け解析とルールベースの情報抽出を組み合わせることにより、特許公報から発明の対象や目的・課題をあらわす情報を抽出[7]

これらの例では、対象分野を限定することにより、単語・複合語や係り受け組では捉える事ができない、深い意味内容を抽出しており、高度な分析が可能となっている。

2.2.4 文書分類

文書分類には、あらかじめ設定されたカテゴリに文書を分類する技術（クラシフィケーション）と、ボトムアップにグループ化を行ってカテゴリを自動生成しながら分類する技術（クラスタリング）がある。前者の分類技術に関しては、従来は人手で分類規則を書くアプローチが主流であったが、最近では機械学習を用いたアプローチ、すなわち分類済みの教師例を用意することにより、分類規則をシステムが学習する方式が主流となってきている。後者の分類技術は、あらかじめ分類体系や分類規則を決める必要がないため、発見的・探索的な分類が可能となっている。これらの文書分類をテキストマイニングの前処理として適用することにより、各文書に付与されたカテゴリ情報を集計・分析の単位とすることが可能である[8]。

3. 特許情報の検索・分析技術

図3は、テキストマイニング技術を応用した特許マイニングシステムの概要である。以下では、ATMS/IR.netやATMS/Analyzerのベースになった、富士通社内向けの特許検索・分析システムを例に、検索・分析の処理と機能について解説する。

3.1. 検索・分析処理の概要

特許マイニングシステムで特許情報の検索・分析を

5）各文書を、単語の重みを要素として持つ多次元ベクトルとして表現し、ベクトルの内積により文書間の関連度・類似度を計算する（ベクトル空間モデル）。
6）斉藤 孝広, 渡部 勇. 障害情報からのマイニング, 情報処理学会 研究会報告. FI-61-20 NL-142-20（2001）
7）田中一成：特許文書の多観点分類について．情報処理学会 研究会報告 NL-161-10，p.69-74（2004）
8）文書分類技術を使うと、例えば、文献情報（論文）に特許分類（IPCなど）を付与することが可能である。

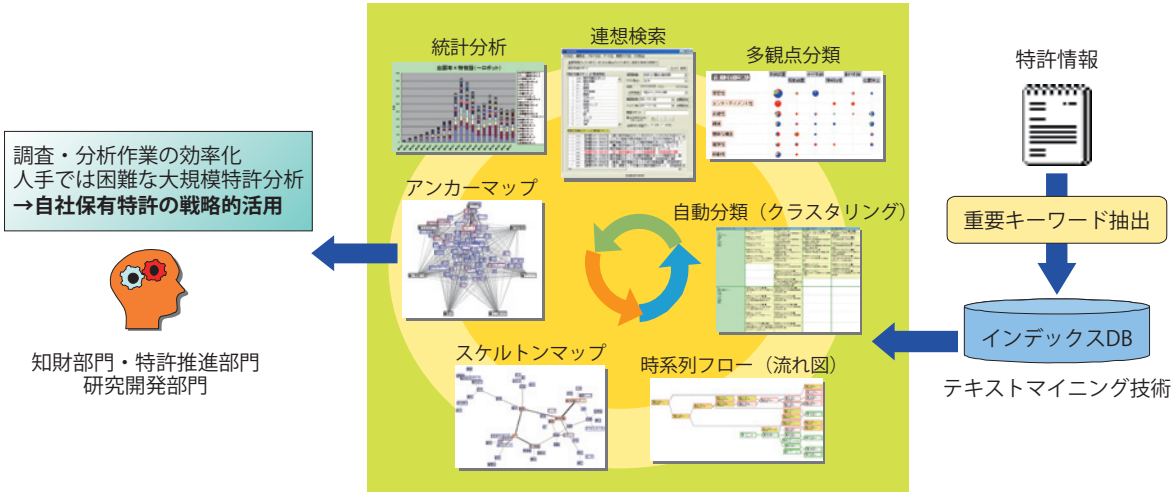


図3　特許マイニングシステムの検索・分析処理の概要

行うためには、まず検索・分析処理で使用するインデックスDBを作成する必要がある。

インデックスDB作成の過程では、特許情報のテキスト部分（名称・要約・請求項・詳細な説明）に対して、単語切出し、頻度集計、複合語構成・分割処理、係り受け解析（主語・述語、修飾語・被修飾語などの単語間の関係を抽出）を行い、キーワードを抽出する。抽出されたキーワードには、統計計算により重要度が付与される。特定の特許にしか出現しない特徴的なキーワードには大きな値が、どの特許にも出現するような一般的なキーワードには小さな値が設定されることになる。上記の処理により、特許ごとに重要度付きのキーワード群が登録されたインデックスDBが作成される。

なお、出願人（特許を出願した組織名）・出願日・IPC（国際特許分類）・FI（ファイルインデックス）・Fターム（特許分類）などの書誌情報も種別ごとにインデックスDBに登録され、検索・分析に利用することが可能である。

検索・分析時には、インデックスDBを用いて、単語間・特許間の関連度（関連性の強さ）が計算される。単語間の関連度は、単語の共起度（二つの単語が互いに同一特許中に出現する度合い）を用いて計算され、同一特許の中で同時に現れる回数の多い単語ペアほど関連度の値が大きくなる。特許間の関連度は、単語の共有度（二つの特許が同一単語を共有する度合い）を用いて計算され、共通の単語を多く含む特許ペアほど関連度の値が大きくなる。

なお、関連度の値は前処理の段階であらかじめ決まっている固定的なものではなく、検索・分析実行時に計算される動的な値であり、分析対象となる特許群を絞り込むことによって変化していく。

3.2. 検索・分析機能の概要

特許マイニングシステムには、特許情報の検索・分析を支援する以下の機能が実装されている。

(1) 連想検索

通常のキーワード検索機能（キーワードを入力して特許をランキング検索）に加え、「関連単語検索機能」「類似特許検索機能」などがある。検索・分析対象となる特許集合の絞込みに使用する。

(2) 統計分析

検索結果をリアルタイムで集計し、グラフ化する。書誌情報やキーワードの出現傾向の分析（IPCや出願人の経年変化、トレンドキーワードの分析など）に使用する。

(3) アンカーマップ

指定した単語を頂点に、その関連語を多角形の中に配置した概念マップの表示機能であり、単語の位置関係により、頂点に指定した単語間の特徴を表示する。比較分析（例えば出願人ごとの特徴比較など）に使用する。

(4) スケルトンマップ

骨格となる強い単語間関連情報だけを表示した概念マップの表示機能であり、単語間のつながりにより、主要な概

念を表示する。特許集合全体の概要把握に使用する。

(5) 自動分類（クラスタリング）

特許を内容の類似性により自動分類（クラスタリング）し、出願人などの書誌情報によって表形式に整理する。特許集合全体の概要把握、特許集合の絞込みなどに使用する。

(6) 時系列フロー（流れ図）

内容の類似性や、引用・参照関係などを用いて、特許間の時間関係を可視化した流れ図を表示する。技術動向調査、基本特許の発見などに使用する。

(7) 多観点分類

係り受け解析と情報抽出の技術を用いて、特許の目的や対象を抽出する。特許を目的別に分類したり、目的と対象の対応分析を行ったりする際に使用する。

(8) 引用分析

特許の明細書（書誌情報と本文）から、ほかの特許・論文への引用情報を抽出する。時系列フローの基礎情報として利用したり、被引用数（ほかの特許から何回引用されているか）を計算することにより有力特許発掘の基礎情報として利用したりする。

上記八つの機能群は相互に連携しており、ある機能の結果から別の機能を呼び出すことができるようになっている。

3.3. 特許検索における利用シーン

以下では、特許の効率的な検索を可能にする関連単語検索機能、文章検索機能・類似特許検索機能を、利用シーンに沿って紹介する。

3.3.1 関連単語検索機能

特許検索においては、適切な検索式を組み立てる（あるいはキーワードを入力する）必要がある。検索結果が粗すぎると内容チェックにコスト・時間がかかり、逆に絞り込みすぎると検索漏れが出てしまう可能性がある。特許検索のエキスパートは、同義語や特許分類（IPC・FI・Fタームなど）を活用することによって、検索効率を高めているが、一般の研究者・技術者にとっては効率的な検索を行うことは容易ではない。ここで紹介する「関連単語検索機能」は、対象技術分野に関する同義語や特許分類の発見を支援し、効率的な検索を行う。

関連単語検索機能による検索例を図4に示す。画面の最上段は検索キーワードを入力する領域であり、中段左側には入力単語に対する関連単語が、下段には入力単語を含む特許がランキング表示されている。関連単語の表示領域には、左側の図では「アーム」の関連単語が、中央の図では「アーム」の関連IPCが、右側の図ではIPC「H01L 21/68」の関連単語が、それぞれ表示されている。図4の例のように、アームの「回転」の同義語・類義語として「回動」「旋回」といった単語を見つけたり（左側の図）、また、特定のキーワードに関連したIPCを探し（中央の図）、そのIPCの関連語を調べることによって（右側の図）、IPCの意味を推定したりすることも可能である。関連語としては、インデックスDBに入っているキーワード・書誌情報を種別ごとに表

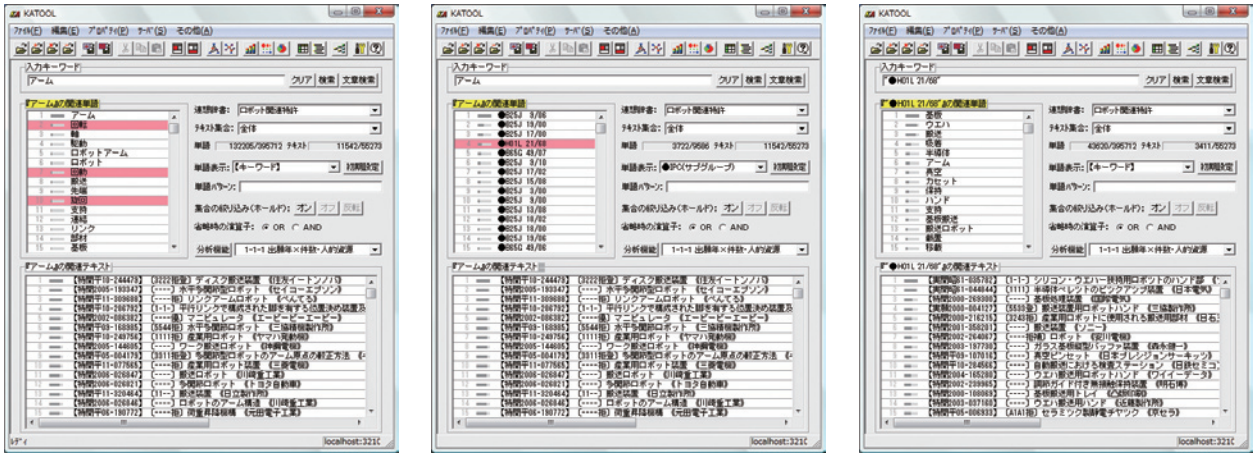


図4　関連単語検索機能による検索例

示することができ、また、特定の文字列パターンにより、表示単語の絞込みを行うことも可能である。

以上のように、関連単語検索により、検索対象を絞り込んだり広げたりするための同義語や特許分類を見つけることができ、特許検索のエキスパートでなくても効率的な検索を実行することが可能となる。

3.3.2 文章検索機能・類似特許検索機能

「文章検索機能」を用いることで、検索式やキーワードを指定する代わりに、文章から関連特許を検索することが可能である。たとえば、特許公報の一部分（特定の請求項など）を指定したり、新聞記事や論文やWebページなど検索対象の特許DBには含まれていない文章を抜き出して指定するといった使い方を想定している。

また、特定の特許を指定して、その特許に内容が類似する特許を検索する「類似特許検索機能」では、調査対象となる特許や、検索中に見つかった関連特許などを入力特許として指定する。

図5は、類似特許検索機能を使用し、ある特許を指定して、その類似特許の検索を行った検索例である。左側の図で番号指定された特許に対する類似特許が、中央の図の下段のリストにランキング表示されている。この類似特許から更に関連がありそうなものをピックアップして（ピンクの網掛けで表示）、その特許群をキーに再度類似特許検索を行ったのが右側の図である。

文書検索や類似特許検索では、検索入力と内容が類似した特許を類似度順にランキングすることが可能であり、キーワードを指定せずに、文章や特許を出発点として、その類似特許を次々と見つけていくことができる[9)]。

3.3.3 そのほかの検索支援機能

特許検索支援機能としては、「関連単語検索機能」「類似特許検索機能」が中心となるが、「自動分類（クラスタリング）」「時系列フロー（流れ図）」などの分析系の機能を、検索の補助に利用することもできる。「自動分類」では、内容の類似性による特許が自動分類されるので、調査対象が含まれる分類を中心に調べていくことで、調査効率を上げることが可能である。また、時系列フローでは、時系列的な関係性が表示されるので、調査対象特許の上流に位置する特許（先願の類似特許・引用特許）を中心に調べていくことで、やはり調査効率を上げることが可能である。

3.4. 特許分析における利用シーン

以下では、技術動向調査などにおいて使用する、特許分析機能を具体的に紹介する。

3.4.1 統計分析

統計分析は、検索結果をリアルタイムで集計して、

図5　類似特許検索機能による検索例

9）文書検索や類似特許検索と、定型項目（特許分類や出願人など）による検索を組み合わせることも可能である。

グラフ化する機能である。グラフの横軸・縦軸には、書誌情報とキーワードを自由に組み合わせて指定することが可能であり、出願年×出願人（出願人の経年変化）、出願年×キーワード（トレンドキーワード）、出願人×キーワード（出願人ごとの特徴キーワード）、出願人×出願人（共同出願人の分析）など、様々なグラフを作成することができる。グラフの縦軸の計算に、分布の偏りを表す統計量を利用することにより、変化がある部分を強調して表示する特徴量グラフを作成することも可能である。

図6は、ロボット関連特許（約3万件の集合）に対して、横軸に出願年を、縦軸に「～ロボット」という文字列パターンのキーワード（「ロボット」で終わるキーワード）を指定して作成したトレンドキーワードグラフである。グラフからは、ここ数年の傾向として「移動する手段を持ったロボット」の特許が増加傾向にあることを容易に読み取ることができる。

統計分析は特許分析の基本機能であり、まず全体としての特徴・傾向・変化などを概略としてとらえるために使用する。詳細な分析は、次節以降に説明する諸機能を用いて行う。

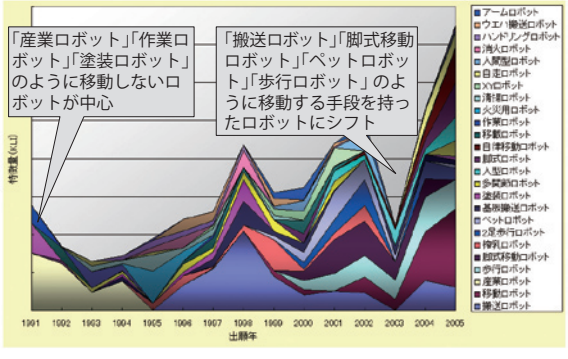


図6　トレンドキーワードグラフ

3.4.2 アンカーマップ

移動ロボット関連特許の集合に対して作成したアンカーマップを図7に示す。図では、出願人の上位9社がアンカーとして9角形の頂点の位置に固定されており、9角形の内部には移動ロボット関連のキーワードが配置されている。9角形内のキーワードは、各頂点から単語間の関連度に応じた力で引っ張られており、その位置は各頂点からの引っ張り力のバランスによって決まる。したがって「ロボット」や「センサ」といった移動ロボットの共通キーワードは真中に、各出願人を特徴付けるキーワードは各頂点の近くに配置される。このようにアンカーマップでは単語の位置関係を見ることによって、出願人ごとの特徴比較を直感的に行うことが可能である。

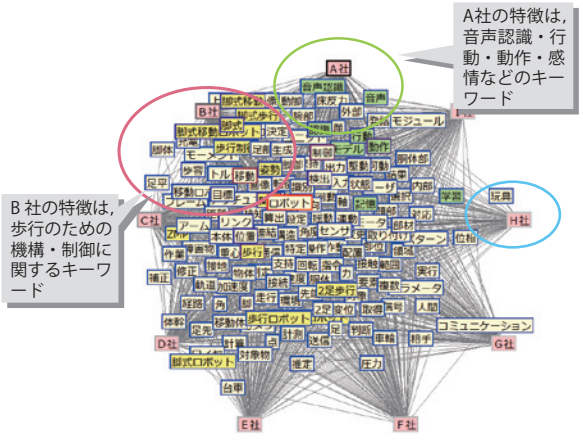


図7　アンカーマップ

3.4.3 スケルトンマップ

図8は、移動ロボット特許に頻出するFI（サブグループ）とキーワードの間の関連性を表したスケルトンマップである。スケルトンマップでは、関連度が小さい関係を削除することにより、骨格となる構造（主要な関係）を表示する。中心的なテーマとなる重要な情報が、放射状の中心（ハブ）になる傾向があり、このハブを順に見ていくことで、全体の概観を把握することができる。

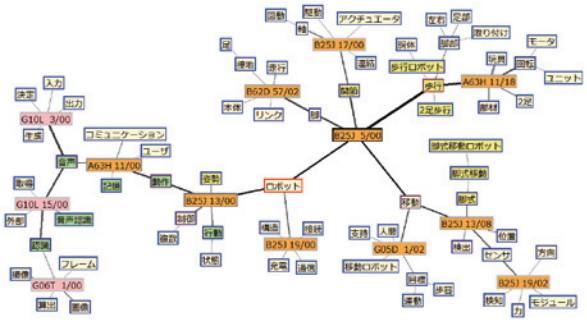


図8　スケルトンマップ

3.4.4 多観点分類

図9は、歩行ロボット関連特許の集合に対して作成した多観点分類のグラフである。横軸には出願年、縦軸には各特許から抽出された特許の目的・課題が表示されている。グリッド上に配置された円は、特許の出願件数を表しており、出願人によって色分けされている。

図からは、1992 ～ 1998年にかけては、B社が中心となり「安定性」「自由度」「精度」といった、歩行ロボットが有すべき基本的な性質に関する特許が多く出願されていたことが分かる。また1999 ～ 2004年にかけては、メインプレーヤがA社に代わり、「安全性」「軽量化」「小型化」「エンターテインメント性」「自律性」といった、歩行ロボットが家庭に入ったときに求められる高度な性質に関する特許が多く出願されていることが分かる。

このように、多観点分類を用いることにより、特許分類やキーワードのグラフからは得られない詳細なトレンド・特徴をとらえることが可能であり、直感的に解釈しやすい結果を得ることができる。

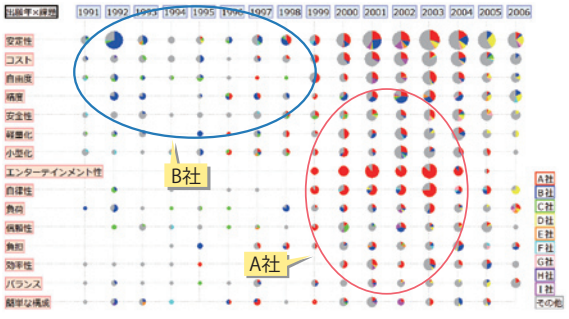


図9　多観点分類

3.4.5 そのほかの分析支援機能

技術動向調査を行う場合には、このほかに「自動分類（クラスタリング）」「時系列フロー（流れ図）」などの機能を使うことができる。「自動分類（クラスタリング）」は、特許群を、内容の類似性を用いて階層的に分類する機能であり、特許群の全体概要を俯瞰（ふかん）したり、人手で付与された特許分類（IPC、FI、Fタームなど）とは異なった観点で分析したりする際に有効である。「流れ図」は特許間の類似性・引用関係を時系列的に整理した図解であり、技術の流れを調べたり、基本特許・周辺特許の判断を行ったりする際の基礎情報として利用できる。

4. おわりに（今後の予定）

前述した技術を採用した（一部機能除く）特許分析システムATMS/Analyzerは、2007年12月に製品リリースし、多くのお客様にご利用いただいている。2008年4月には、審査経過情報を活用して、客観的に特許の価値評価を行うレイティング機能を追加、2008年10月には、外国語の重要単語（課題や目的など）をフレーズで抽出する技術を発表している。これにより日本の特許情報だけでなく、外国特許、学術文献などの分析も可能となった。

更に、今後は特許管理システムATMS/PM2000で持つ社内情報などもATMS/Analyzerに取り込むことで、特許ポートフォリオ分析もより一層容易になるであろう。

profile

渡部　勇（わたなべ　いさむ）

1985年　慶應義塾大学理工学部機械工学科卒業。
1987年　東京工業大学大学院制御工学専攻修士課程終了。
同年、富士通株式会社入社。現在、株式会社富士通研究所ソフトウェア＆ソリューション研究所ソリューションテクノロジ研究部部長。

profile

待井　学（まちい　まなぶ）

平成3年　（株）富士通長野システムエンジニアリング入社
平成6年　特許ビジネス（ATMS）担当
平成17年　富士通株式会社　ATMSビジネス部へ出向
平成20年　復職　現職