第Ⅳ章　研究者のための特許文献検索システムとは

特許情報を活用するためには、特許文献を検索するためのコンピュータを使って膨大な情報の中から必要とする情報（文献）を取り出すことが第一歩である。より多くの研究者がこの一歩を踏み出すには日常の論文情報の検索で行っているようなキーワード（技術用語）を用いて、インターネットのように「容易」に、そしてインデックス検索のように「精度良く」特許情報を検索することができる道具（検索ツール）が求められる。

本委員会が大学等の研究者に実施したアンケート調査や本委員会における有識者の議論を踏まえ、研究者に対して特許情報の活用を促すために参考となる特許文献検索システムのあり方を以下のようにまとめた。

１．特許情報の検索

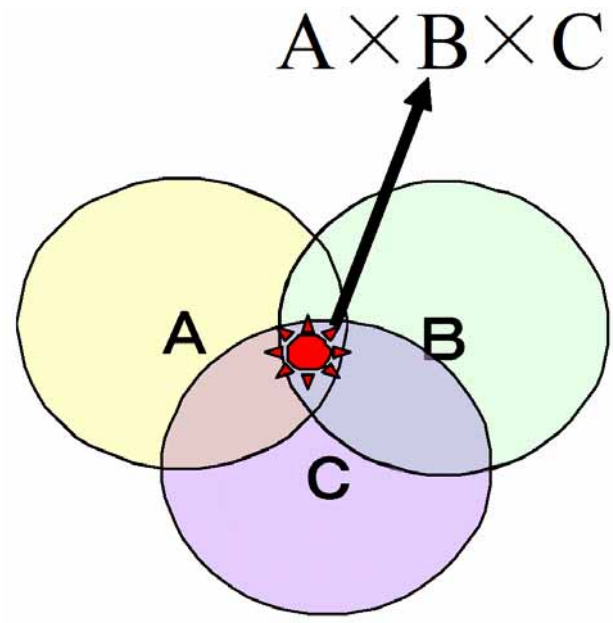
１．１　発明のとらえ方

ここに要素Ａと要素Ｂと要素Ｃからなる発明（調査の対象とする技術と考えてもよい）があったとすると、この発明に対する従来技術を調査するには、それぞれの積集合（Ａ×Ｂ×Ｃ）で検索式をたてるのが一般的である。この方法は、論文に関する調査をする場合にも基本的には同じである。

ところが、特許を出願する際や審査請求をする際に行う調査は上述の調査だけでは不十分である。すなわち、Ａ×Ｂ×Ｃの検索式で必要な文献を発見できなかった場合には、Ａ×Ｂ、Ｂ×Ｃ、Ｃ×Ａのそれぞれの検索式でも調査をして、発見した複数の文献を組み合わせることにより要素Ａ，Ｂ，Ｃからなる発明を構成できないかどうかを検討する必要がある。実務上は、最初にＡ×Ｂ×Ｃの検索式で調査を行い、全ての要素を開示した文献を発見した場合にはその時点で調査を終え

ることになるが、通常はそのような文献は発見できないので、発見した中で最も技術的に近い文献と調査対象としている発明の相違点をＡ×Ｂ、Ｂ×Ｃ、Ｃ×Ａなどの検索式で探すことになる。

図 Ⅳ-1 発明(技術)＝Ａ×Ｂ×Ｃの調査



１．２　インデックス検索

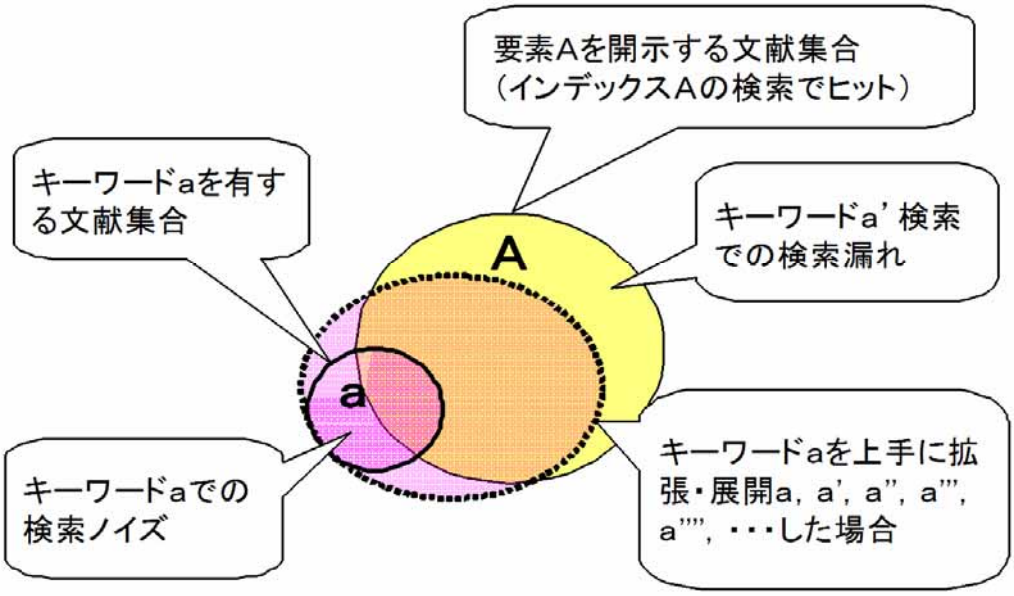
ＦＩやＦタームなどのインデックスを用いた検索では、図 Ⅳ-1 のＡ，Ｂ，Ｃそれぞれに該当するインデックスの中に原則として該当する技術のすべての文献が入っているので上記の調査で必要な文献は網羅することができる。しかし、１９万インデックスのＦＩや３５万インデックスのＦタームを使いこなすことは容易なことではない。

１．３　キーワード検索

そこで、一般にはキーワードを用いた検索がよく行われる。ところが、キーワードを用いて十分な調査をするには同義語や関連語などを考慮して用語の拡張・展開を行いながら当該キーワードの周囲も含め

て調査をする必要がある。例えば、次の図のようにインデックスＡ（例：内燃機関）で検索できる技術の文献集合をキーワードで検索するには、［ａ＝内燃機関］というキーワードだけではなく［ａ’＝ガソリンエンジン＋ガソリン機関］、［ａ''＝ディーゼルエンジン＋ディーゼル機関］、［ａ'''＝ジェットエンジン＋ジェット機関］、［ａ''''＝ロータリーエンジン＋ロータリー機関］などの用語に拡張して検索を行う必要がある。この場合エンジンに引きずられて「検索エンジン」や「サーチエンジン」もノイズとして抽出してしまう可能性がある。

図 IV-2 キーワード検索の概念



さらに、キーワードを用いた調査では本来調査すべき範囲を網羅するために何通りかの検索を繰り返すことになるが、検索の度に検索漏れや検索ノイズが含まれる可能性があり、これを検索者の「技」だけに任せると人によって様々な調査結果が生じることになる。

この点に関して、特許庁が行った調査では「検索性能を高めるためには、“検索漏れをなくす”、“検索ノイズを少なくする”の２つの側面からの問題解決に当たる必要がある。一般に文書検索では、検索対象となる文書の表現（文書の内容を表す索引語等）と検索質問の表現（検

索者の情報要求の内容を表す特徴語集合）との照合によって検索が行われる。文書の索引語表現と検索質問の表現との間にミスマッチが生じると、検索すべき内容の文書が検索されない（検索漏れ)、検索すべきでない内容の文書が検索されてしまう（検索ノイズ）などの問題が生じる可能性がある。」(「次世代特許審査システム用検索ツールの基礎調査」特許庁平成１６年３月）として、次のような課題をあげている。

①表記の揺れの存在
②同義表現への対応
③キーワード（特徴語）の選択性
④語義の多様性
⑤複合語への対応
⑥概念レベルでの照合
⑦文の構造を反映した言語表現レベルでの照合

１．４　概念検索

概念検索は、多くの検索ツールで実用に供されている。そして、自らが欲する情報を文章にして入力すると瞬時にしてその回答を類似度順に表示するようになっている。ところが、「概念」とはいっても、結局、コンピュータ技術者が設定したロジックに沿ってプログラムを作動させているので、現状のキーワード検索が有する検索漏れやノイズの問題はそのまま内在している。
概念検索について、特許庁が行った検索ツールに関する調査では次のように言及されている。

- 「現時点では単に概念検索のみを利用した検索には大きな期待はできず…」(「次世代特許審査システム用ツールの検証調査」特許庁：平成１４年３月）
- 「概念検索のようなツールは、システムがブラックボックス化しているため…特許実体審査業務のサーチには適さない」(「次世代特許

審査システム用検索ツールの基礎調査」特許庁：平成１６年３月）

すなわち、こうした概念検索のエンジンは、検索アルゴリズムが開示されないため、検索者が抽出されるはずであると考える文献と、実際にコンピュータが抽出の対象としている文献との間にズレが生じやすく、どこまで検索をすれば自らが想定した範囲を調査したことになるのかが理解し難いという欠点がある。

２．研究者に必要な特許文献検索システム

２．１　基本となる情報検索ツール

大学等の研究者が特許情報を活用するための情報検索ツールは、検索精度が高く使いやすいものがよいことは当然であるが、加えて次の点も考慮すべきである。

- すでに活用されている論文検索との連携がとれるものであること
- 論文情報を検索するキーワードでシームレスに特許情報も検索することが可能なこと（論文情報を検索をすると特許情報も違和感なく同時に表示されることが望ましい）
- 検索ツールの情報が公開されており、多くの研究者が知恵を出し合って持続的に機能の高度化ができること
- ユーザーのニーズを機動的かつ柔軟に取り込むことができるものであること
- 大学等において誰でも自由に使えること

以上の条件を満たすものとしては、汎用連想計算エンジンＧＥＴＡが選択肢の一つになるが、条件に合えばほかのツールでも何ら差し支えない。特許の検索の精度を上げることは難しい課題も多いが、上記のような柔軟性のあるツールであれば日本の情報検索の研究者がＬＩＮＵＸのように改良を続けていくことも十分に期待できる。

２．２　検索に関する知的資産の活用

特許情報を検索するツールは、検索式を入力する入り口側の技術から検索結果を検索者に表示するまでに多くの技術的要素があり、これまで多くの研究者が精度の向上や利用性の向上に取り組んでいる。これらの研究成果や、特許庁がこれまで実施してきた検索に関する調査研究の結果を集約して高機能の検索ツールの実現に役立てることが肝要である。また、国際特許分類改正の動向やＦＩ、Ｆタームのリフォ

ームの情報など検索に関わる情報やデータは早期に取得して機動的かつ迅速にシステム変更やデータ修正などを行うことも重要である。

研究開発の二重投資を回避し、効率よく検索ツールの機能や精度を向上させるためには、特定の技術に頼るのではなく、研究者が保有する有用な「特許技術」を必要に応じて導入することも重要である。この点では研究者に必要な特許文献検索システムは特定企業の技術に委ねるのではなく、検索ツールの要素技術毎に優れた研究者の成果を結集することが望まれる。

３．特許文献検索システムの具体的イメージ

３．１　検索の入り口の容易化（検索式作成支援）

研究者は、調査を行う技術分野に関する用語を熟知している。そうすると、概念検索のように自然文で検索式を作成するのではなく、ポイントとなるキーワードを入力した方が必要とする特許情報の輪郭が明確になるとともに、検索の精度向上に必要な技術の関連づけなどの指定がしやすい。

研究者が、特許情報を検索する入り口となる検索式の入力を容易にするため、次のような機能や支援ツールが必要である。

（１）ポイントとなるキーワードの関連語などを拡張して入力するための支援ツール；関連語の拡張には検索結果に含まれる用語を検索者に提示することにより、検索者が用語を連想したり追加的に用語を取捨選択したりできるようにすることも有効

例 1：内燃機関→	**2 サイクル、4 ストローク、エンジン、レシプロ、往復機関、ピストン機関…**
例 2：マネー→	**金、現金、通貨…**

（２）キーワード検索の検索精度を上げるためにキーワード間の意

味的な関係（用語間の係り受け関係：以下の例のように「４サイクルエンジン」と「傾斜」は１文献の中にその双方の単語が含まれているだけではなく、双方の単語が関係を持った記載があること）を持たせて検索式の作成を可能にすること

> **例：[4サイクルエンジン←（係り受け）→傾斜]×小型船**

（３）異表記（「コンピュータ」と「コンピューター」や用語の交差（例：「特許の出願」と「出願の特許」）の処理など、自然言語処理技術で一定の精度が期待できることは自動化すること

（４）利用者固有の技術分野に応じた専門用語登録と不要語登録を可能とする辞書機能

> **例1：「飲食物を入れる容器の構造」の発明で「回転鮨」（「飲食物」という用語を多用する技術分野）の文献が多数ヒット→排除して支障がない場合「不要語」に登録して検索結果から除外**
>
> **例2：「携帯電話の数字3段×3段を用いた日本語入力」の発明で「ワープロの日本語入力」の文献が多数ヒット→排除して支障がない場合「不要語処理」**

上記（２）に関連して、係り受け等で表現される意味的な構造は情報検索にはあまり役立たないというのが定説だった。しかし、この定説はワンショットの検索を前提としており、実はあまり意味がない。というのは、意味的な構造が必要になるような難しい検索がキーワードの単純な組合せによるワンショット検索で完了することはほとんどあり得ないからである。難しいことを人に尋ねるのに２つ３つのキーワードの組合せで済むはずもなく、対話が必要である。意味構造を用いる検索が意味をなすのは対話的に検索質問を改訂するインタラクティブな検索の場合なのである。

さらに言えば、意味構造はそのような場合の検索式のインタラクティブな改訂の支援にこそ有用である。たとえば、「ロボットで家を作る」という自然言語文を検索質問とする場合、「家」と「作る」の間の関係

に基づき、「作る」の類義語として「建てる」、「建設」、「建築」などを優先度を自動的に高めて利用者に提示したりすることが可能であり、これによってインタラクティブな検索の効率を高めることができる。係り受けや照応・共参照で表現されるこうした意味的な関係を求める自動解析の精度は現在のところ７０％に満たないが、そのような自動解析で求めた品質の悪い意味構造を用いた場合でも、それに基づくこのような支援によってインタラクティブな検索の効率が大幅に向上することが明らかになりつつある。また、文書等のコンテンツの作成の段階からその意味構造を人手によって明示しておくこと(セマンティックオーサリング)により、実はコンテンツの作成が楽になり、なおかつコンテンツの品質が高くなることもわかっている。しかも、そのようにして作られたコンテンツは上記のような意味構造に基づくインタラクティブな質問改訂を通じてきわめて効率的に検索することができるはずである。このように考えると、検索のみならず特許文書のライフサイクル全体を通じた知的生産性の向上を図るには、セマンティックオーサリングに基づくコンテンツの作成が重要であると言えよう。

このような観点を敷衍すれば、検索だけを最適化しようとするのではなく、技術に関する知識の生成と循環と拡大再生産のサイクル全体にわたる最適化を考えるべきであり、セマンティックオーサリングはその意味での全体最適化をもたらすものと期待される。

３．２　検索ツール（エンジン）の精度向上

キーワード検索でも検索漏れや検索ノイズを回避してインデックス検索に近い精度を確保するためには、少なくとも特許庁が行った「次世代特許審査システム用検索ツールの基礎調査」（平成１６年３月）に挙げられた課題（上記１．３①～⑦）への対応が必要である。対応策の具体例を挙げるとつぎのとおりである。

（１）係り受け関係を考慮したキーワードとのマッチング処理（上記３．１（２）で検索式を反映するための処理）
（２）表記の揺れ（例:「コンピュータ」と「電子計算機」)、異表記、用語の交差、類似度の処理（不完全一致の処理）
（３）単位を伴う数値（例：５ｃｍ以上）条件の処理
（４）部分語の識別可能化(スキーでウイスキー､帯電で携帯電話、Ｈ２でＨ２Ｏがヒットするのを排除)
（５）否定語の処理:「除く」､「含まない」などの識別

さらに、特許情報を対象とする検索に特有の処理として、国際特許分類（ＩＰＣ）の活用が考えられる。たとえば、キーワード検索でヒットした一次文献のうち上位１００件程度を統計処理して、付与されたＩＰＣサブクラス（全体で約６５０）またはＦタームのテーマ（全体で約２，６００）の上位２～３（１件の特許情報の副テーマ率を考慮）を用いて絞り込み、キーワード検索のノイズを防止することが考えられる。このようにＩＰＣを利用する場合、研究者が表示された候補を選択するか、自動的にＩＰＣを適用して絞り込むことが考えられる。

また、論文情報と特許情報の用語の相違をつなぐような連携辞書、特許に特有の技術用語を扱う処理、各情報に含まれる引用文献情報の自動抽出なども利便性を上げるうえで重要である。

こうした課題を実現する検索ツールの一つの候補として上述の汎用連想計算エンジン（ＧＥＴＡ）があげられる。

さらに、おもなユーザーとして研究者を想定する検索ツールでは、論文情報から関連する特許を検索する機能が特に有効であると考えられる。研究者は自分の研究分野の論文に精通しており、日常的に各種の論文検索システムを使いこなしている。関連する研究動向や技術動向についても、重要な論文やサーベイ論文などを通じて把握している場合が多い。この知識をそのまま活用して、関連する特許を検索できれば、研究者が特許情報に日頃から接する有効な手段となる。

論文と特許を結びつける情報として、まず直接的な参照情報を用いる方法が考えられるが、関連論文を網羅的に参照している特許は少なく、その逆に特許を参照している論文は極端に少ない。したがって、論文と特許の間の有用な結びつきを準備するには、何らかの情報技術によって、両者の扱っている技術内容の類似性・関連性を機械的に抽出する必要がある。

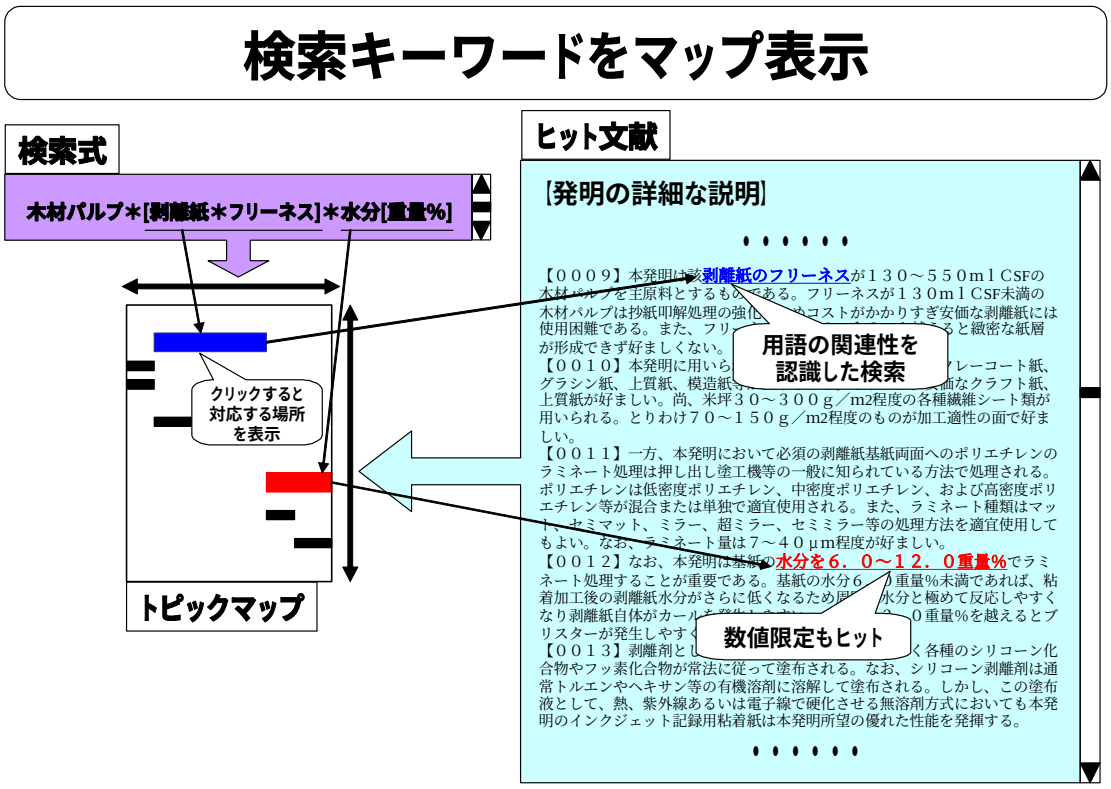
こうした課題を解決する検索ツールの一つの候補として上述の汎用連想計算エンジン（ＧＥＴＡ）があげられる。使用されている言葉の重なりだけを手がかりに文書内容の類似性を評価する手法であるため、論文と特許の用語の違いが問題となりうるが、サーベイ論文や特許では、解決すべき課題の説明や研究開発の経緯について一般的な言葉で解説されている場合も多く、それらが両者を結びつける役割を果たすと期待できる。ただし、より精度の高い関連づけのためには、論文や特許で使用されている基本概念や用語の整理とそれらを関連づけるオントロジや辞書などの整備が必要であると考えられる。

３．３　検索結果の易読化

学術論文と同じく特許文献（明細書）は文章が長いため、検索した各特許文献が検索目的に適合するかどうかを判定するのに多くの時間を要してしまう。また、特許文献を理解する際に請求項の理解は避けて通れないが、請求項読解は特許に慣れていない研究者にとって困難を伴う。これらの問題を解決する方法が、東京工業大学（岩山客員助教授）で研究されている。

特許文献のような長い文書の読解を支援するためには、文書内のトピック分布を可視化して長い文書中の目的箇所へ容易に到達できるようなインターフェイスを提供することが必要である。これは目次や索引を自動的に作ることに相当する。東京工業大学では、図 IV-3 に示すような「トピックマップ」と呼ばれる文献読解インターフェイスを提案している。トピックマップの活用により、特許検索プロセスに占める特許読解の負担を減らすことができる。

図 IV-3 トピックマップ(検索キーワード)



３．３．１　検索キーワードとヒット文献のマップ化

トピックマップは、検索式の各キーワードがヒット文献内のどこに現れているのかを二次元のマップで図示するものである。マップでは、横軸が検索式に相当し、縦軸がヒット文献に相当する。つまり、検索式中の最初の文字がマップの左端に、最後の文字が右端に対応する。同様に、ヒット文献中の最初の文字がマップの上端に、最後の文字が下端に対応する。トピックマップ中の矩形は、その場所に、検索式、ヒット文献で重複する部分があることを表している。矩形の、縦横それぞれの座標に相当する文字列が対応していることになる。よって、検索式中の任意のキーワードを指定すると、トピックマップを介して指定キーワードに対応するヒット文献中の場所を直接表示することができる。このことにより、ヒット文献中の関連部分を飛ばし読みする

ことができ、ヒット文献が検索対象に適合しているか否かを効率良く判定することができる。

検索式とヒット文献で対応する部分を抽出する際は、複数の近接する対応をまとめたり、部分的にしか文字列が重複していない対応を抽出したりできる。図 IV-3 の例の場合、ヒット文献中に「剥離紙」と「フリーネス」が近接してあらわれるため、トピックマップでも近接する２つのキーワードをまとめて対応付けている。このことにより、用語の関連性を見つけることができる。同様に、ヒット文献中の「水分を６．０～１２．０質量%」というフレーズが「水分」と「質量%」という２つの検索キーワードと対応付けられているため、「６．０～１２．０」という数値限定を容易に見つけることができる。他にも、「コンピューター」と「コンピユーター」、「特許検索」と「特許を検索（する）」などの表記ゆれも自動的に対応付けることができる。

３．３．２　特許文献内の対応関係のマップ化

トピックマップは、特許文献の読解支援にも使うことができる。特許文献は、請求項を中心に読む場合が多い。ところが、請求項は独特の文体で記述されるため、特許に慣れていない研究者にとっては請求項読解に多くの時間がかかってしまう。請求項の内容を把握するには、請求項だけを読んでいたのでは不十分で、請求項に対応する本文（例えば、実施の形態）も読まねばならない場合が多い。前述したように、特許文献は文章が長いため、請求項に対応する部分を特許文献の先頭から探していたのでは時間がかかってしまう。

トピックマップでは、検索式として請求項全文を指定することで、請求項の任意の部分文字列に対応する本文の場所を一覧することができる。図 IV-4 に例を示す。ここでは、横軸が請求項、縦軸が明細書に相当し、複数の矩形が集中している箇所は両者が強く関連している部分となる。トピックマップの任意の場所をクリックすることで、明細書の対応する部分を頭出しすることができる。

トピックマップは、２つの文書中の全ての部分文字列間で対応が抽出できる。ユーザーは、単に調べたい文章を入力するだけで、その文章内の任意の文字列が検索対象文書のどこに出現するのかを調べることができる。例えば、「関連用語がユーザーが期待した用語でなかった場合」という文章を入力すると、「関連用語」「ユーザー」「期待」「ユーザーが期待」「期待した用語」「ユーザーが期待した用語」等の全ての可能性について、検索対象文書中の出現箇所を数え上げてマップに表示する。

図 IV-4 トピックマップ(用語対応関係)

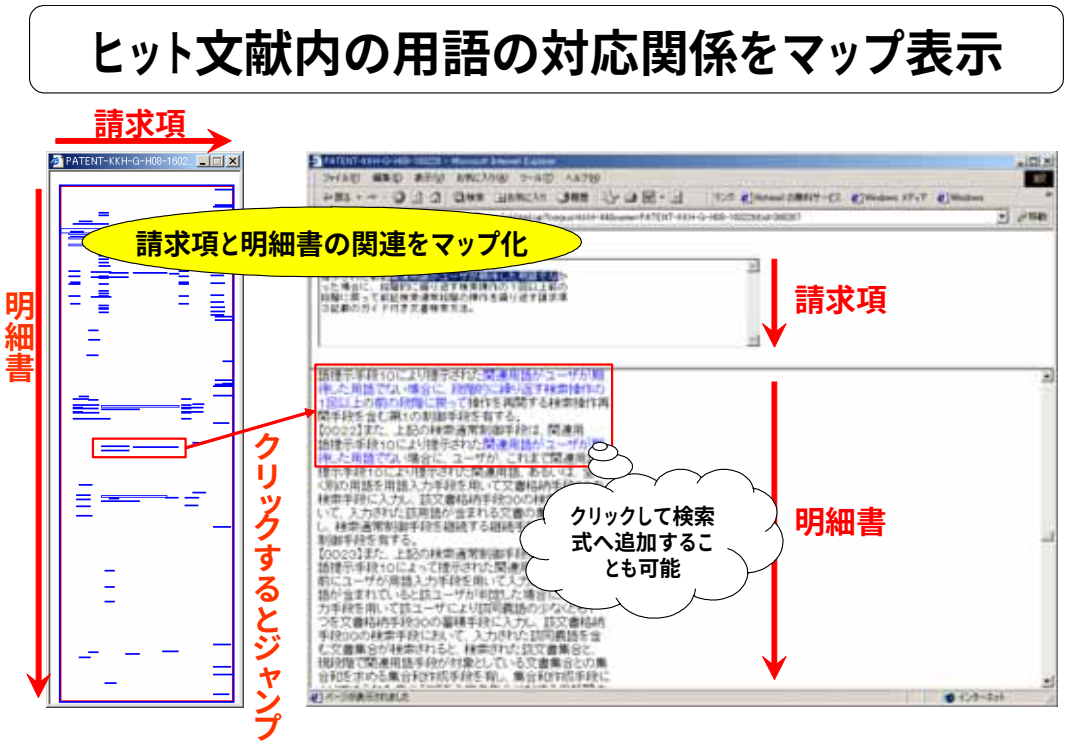
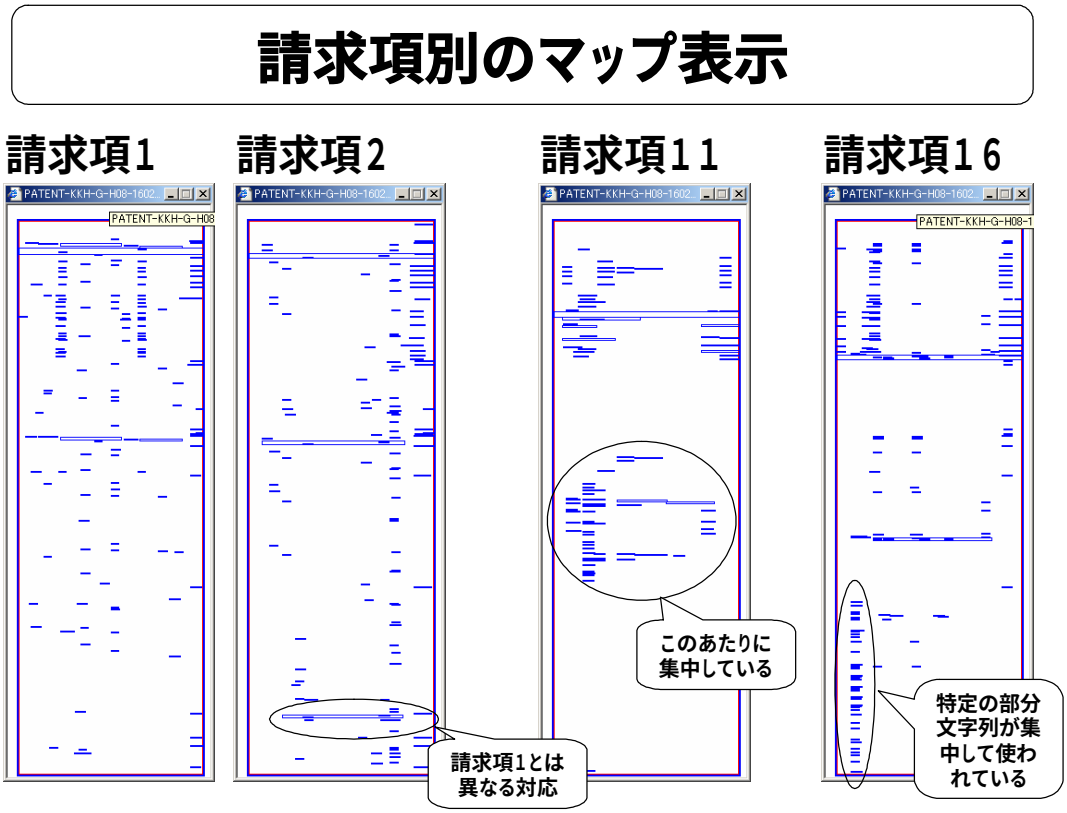


図 IV-5 は同じ特許の異なる請求項毎にトピックマップを表示した例である。請求毎に異なるマップが描かれる点に注目されたい。トピックマップは請求項単位で明細書の目次を作ることに相当する。

図 IV-5 トピックマップ(請求項別)



請求項１と請求項２のトピックマップを比べると、矩形の分布はほぼ同じだが、請求項２の後半に請求項１とは異なる対応を見ることができる。よって、請求項２のみに関連する部分として、この部分に対応する本文を読めばよいことがわかる。また、請求項１１ではマップの中心部分にのみ矩形がまとまって分布している。つまり、請求項１１に関連する本文は明細書の中盤にまとまって記述されていることがわかる。最後に請求項１６のマップを見ると、左下に縦長の分布を見ることができる。これは、請求項１６の前半部の単語が明細書の後半に集中して使われていることを意味している。

３．３．３　特許と特許との関係のマップ化

関連する２つの特許間で、重複する部分、異なる部分を分析したい場合も、トピックマップを使うことができる。この場合、検索式に一

方の特許明細書全文を指定すれば、２つの特許間で重複する部分を一覧することができる。

３．３．４　　ヒット文献表示の際の類似度によるランク付け

検索結果のスクリーニング効率を上げるには、ヒット文献を信頼性の高い方法でランク付けすることも重要である。正解文献（技術的な類似度が高い文献）のみを上位にランク付けすることが出来れば、不要な文献を調べる必要がなくなる。

ランク付けの一般的な方法では、ＴＦＩＤＦと呼ばれる尺度で用語を重み付けし、この重みに従って各ヒット文献のスコアを算出する。複数のヒット文献を比べると、検索キーワードをより多く含む文献ほどその検索キーワードに深く関係する文献と言える（ＴＦ)。一方、複数の検索キーワードを比べると、文献集合全体に渡って出現するキーワードより、少数の文献にのみ出現するキーワードの方が絞込みには有用である(ＩＤＦ)。ＴＦＩＤＦでは、上記２つの尺度により用語を重み付けする。また、用語の重みから文献のスコアを算出する際は、長い文献と短い文献を平等に扱うように文書長の正規化を行う。ランク付けには様々な方法が提案されているが、いずれもランキングは目安にすぎず､「ここから下位は調べなくてもよい」といった明確な基準を提供してくれるわけではない。

国立情報学研究所（ＮＩＩ）が主催する評価ワークショップ「ＮＴＣＩＲ特許検索タスク」では、特許明細書に特化したランク付け手法をコンテスト形式で比較評価している。そこでは、まずコンテストの参加者を募り、共通の検索課題と検索対象を参加者に与える。参加者は与えられた検索課題に対し検索を行い、検索結果を提出する。最後に主催者が検索結果を比較評価する。共通の環境で特許検索技術を比較検討することで、技術の底上げや技術進展の加速をねらっている。ＮＴＣＩＲ特許検索タスクは過去３回実施され、明細書の構造を加味したランク付け手法や、明細書に固有の用語重み付けの手法などが、

参加者から提案されている。

３．４　特許文献検索システムの具体的機能のまとめ

以上で述べた大学等の研究者が使いやすい特許文献検索システムの具体的機能をまとめると以下のとおりである。

- 検索式は、キーワードから出発する。その関連語の拡張支援ツール、検索結果から用語を連想・追加するツールを装備。
- 入力したキーワード間の関連性を持たせて検索式を作成可能。
- 自然言語処理で一定の精度が期待できることは自動化。
- 利用者固有の技術分野に応じた専門用語登録と不要語登録を可能とする辞書機能を装備。
- 数値条件の処理や否定語の処理が可能。
- 国際特許分類（ＩＰＣ）も自動又は手動で活用できる支援ツールを装備。
- 論文情報と特許情報の用語の相違をつなぐような連携辞書を装備。
- 文献に含まれる引用文献情報は自動的に抽出可能。
- 検索キーワードとヒット文献の関係、特許文献内の対応関係などはマップ化して表示可能。
- ヒット文献は類似度によりランク付けして表示、など。