「発明」2003年４月号　(hatsu0304)

概念検索による特許情報の活用法　その１

六車正道　（むぐるま　まさみち：㈱日立製作所　知的財産権本部　特許主幹，技術士）

要約　技術的な説明文を質問として入力するだけで，類似度の高い順に特許がリストアップされる概念検索システムが実用になってきた。概念検索では，同義語の追加はほとんど不要であり，検索に不慣れな研究者・特許技術者でも簡単に，短時間で利用することができる。しかし，概念検索は発展途上であり，いろんな方式があり，上手な質問文の作成も少々コツがある。さらに，概念検索に使われるワードを参照したり，重み付けを変えたり，また，必須のワードやIPCを指定することで，いっそう効果的な検索が行える。

今回と次回は概念検索の基本的な仕組みや上手な利用法を紹介し，３回目は，公知例調査やアイデア発想支援ツールとしての概念検索の利用を紹介する。

１．はじめに

情報環境の進展は著しく，技術者自身が特許情報の検索を希望することが増えてきた。また，技術開発競争の激化を反映して技術者自身による特許情報の調査の潜在的な必要性も増加している。特許調査は，他社特許の侵害防止や自社特許の権利範囲を確定するための調査だけでなく，他人の特許を参照することで自分のアイデア発想の大きな飛躍に役立つことが期待されている。

特許情報の検索は，特許庁の無料の電子図書館・IPDLのサービス開始もあって一般化してきている。検索は大別して，①特許番号による検索，②国際特許分類・IPCや出願人による検索，③キーワードによる検索に分けることができる。この中で，キーワードによる検索は，IPCのような特殊な知識が不要であり，対象技術者ならば誰でも知っている技術用語で検索でき，また，特許情報を詳細に絞り込むことができるので，重要なものである。しかし，キーワードはIPCのように定義が明確でないために，いくつもの同義語を補充するようなことが必要であったり，IPCとの組合わせなどが必要であった。このため，検索を行なうための検索式の作成は熟練が必要であり，一般技術者にとっては敷居の高いものであった。

これに対し，探したい情報を説明する文章をそのまま質問文としてコンピュータシステムに与える概念検索[1)]が数年前から話題にのぼっている。さらに，概念検索の回答はヒット（該当）の可能性の高い方からリストアップされる優先順回答方式である。「概念」とは，ある事物の概括的な意味内容ということであり，キーワードがポイントを指すのに対して，広がりのある範囲を指すという意味合いがある。

日本特許を対象とする概念検索は，2000年夏からＮＲＩサイバーパテントデスク[2)]において，要約とクレーム（権利請求の範囲）を対象として商用サービスが始まった。また，2001年夏からは日立製作所において，平成５年以降の明細書全文を蓄積して社内で大規模な実用化が開始[3)]された。しかし，一般的には概念検索は再現率が低く，実用にならないという主張もあり，情報検索が実用になるとの決定的な評価はまだ得ていないようである。

概念検索の考え方自体は古くからあるが，膨大な日本特許を対象に実務に使えるものはこれまで存在しなかった。筆者は約２年にわたり，概念検索の社内での利用促進の活動を行なうと共に，機能改善を推進してきた。この結果，社内においては十分実用になるとの地位を確立し，研究者や特許技術者などのエンドユーザを初めとして広く利用されるに至っている。また，寄せられた質問への回答が整備され，活用ノウハウも蓄積されてきた。そこで，本稿では，特許情報の活用において，概念検索がすでに実

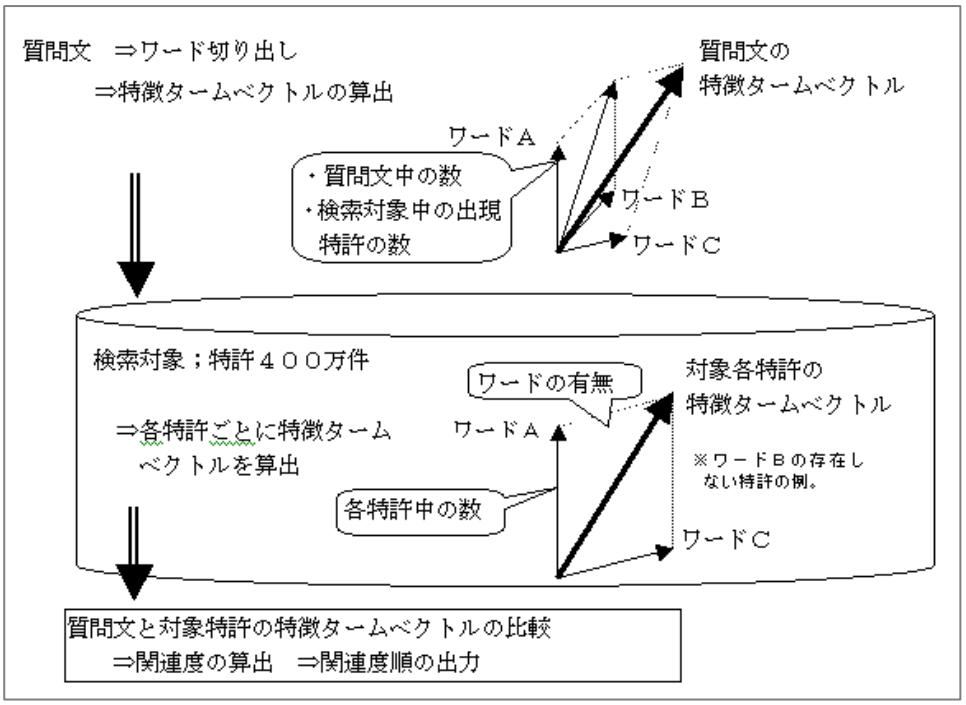

図１　概念検索システムの原理

用化に至っていることを，実例を交えて紹介する。

２．概念検索とは何か？

概念検索システムの明確な定義はないようなので，ここでは，検索条件として文章，またはいくつかのワードを与え，回答は内容的にヒットした可能性の高い順（優先順）に出力されるものとする。

なお，検索式では検索者が重要と判断した「キーワード」を利用するが，概念検索ではコンピュータが「ワード」らしきものを切り出して処理する。例えば「インターネット競り取引で・・・」という質問文では，インターネット，競，取引などともに，「ネット競」などをワードとして切り出すことがある。このため，概念検索で切り出されただけのワードは，本稿ではキーワードとは言わない。

２．１　概念検索の原理

現在の概念検索の基本は，質問文中の各ワードの，対象データにおける出現頻度を基礎に，質問文と対象との関連度を算出するものである。しかし，細部はシステムによって異なり，結果に大きな違いが出てくる。この事実を過小評価して安易に概念検索一般を論じると誤りをおかす恐れがある。図１は，筆者の利用している概念検索システムの原理を示すものである。まず，質問文中のワードの点数を，以下のやり方で決める。

①質問文中に何度も出てくるワードは点数が高い。

②データベース全体でそのワードを含む蓄積文書の少ないワードは点数が高い。

次に，各ワードの点数の合計を算出する。この計算は，各ワードの数値を二乗して和を出してその平方根を求めるベクトル計算である。これにより，質問文にただ1つの大きさと角度の値が得られ，これを「特徴タームベクトル」という。

次に，蓄積された全特許を対象に下記の観点で計算する。

③質問文中の多くのワードを持つ特許は点数が高い。

④各ワードの出現回数の多い特許は点数が高い。

次に，この点数の合計を算出し，さらに質問文の特徴タームベクトルと比較し，各特許の関連度を算出し，点数の高い方から数十件ずつ表示させる。したがって，質問文中のワード全ての and でも or でもない。また，ワードを全て持っている特許が最高の点数とも限らない。

また，検索式方式のように回答件数をいうことはできない。あえて言えば回答件数は切り出されたワードの or 検索の全件であるが，多くの場合数万件以上になり，参照できる上限を超えている。概念検索において，検索できないとか漏れるということは，点数が低かったので見なかったということであり，多くの場合，絶対に見られないものではない。しかし，一定以上の件数を見るのは現実的に困難であるから，利用者が限界点を設けなければならない。

(1) 適合率や再現率の考え方

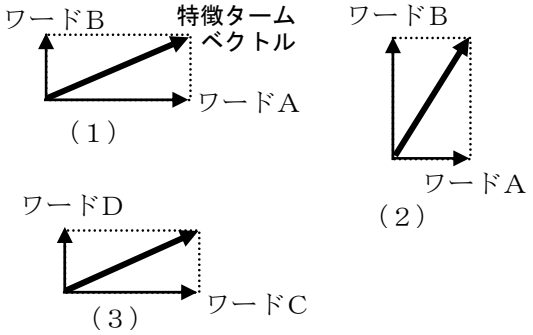

図２　特徴タームベクトルの計算

概念検索では与えたワードが1つでも入っていれば対象になるので，点数の低いものを含めると極めて大量の特許を対象にしている。しかし，実用的には上位 50～100 番目（多くても 200 番目）あたりまでに自分の欲しいものが存在しない場合には，別の質問文を考えて再度検索する。それを 3，4 回繰り返す。その結果得られた類似特許を回答とし，自分が見た全件を母数として，適合率を算出する。再現率は，他の調査で得られている類似特許があればそれらを含めた合計値を母数として，算出する。

(2) 角度は意味があるのか？

図２はワード２個の場合の特徴タームベクトルの計算を図にしたものである。同図(1)と(2)の特徴タームベクトルは同じ大きさである。しかし，(1)は最初のワードＡが大きいのに対し，(2)は二番目のワードＢが大きいという違いがある。この違いを表すために角度が必要になる。

(3) 大きさと角度が同じなら似ているのか？

図２の(3)はワードＣ，Ｄの場合だが，(1)と比べて特徴タームベクトルの大きさと角度がほとんど同じである。このように，大きさと角度は同じだが似たものでないことをどう考えればいいか。これは，ワードＡ，Ｂと，ワードＣ，Ｄという別のフィールドで比較しているわけで，意味のないことである。つまり質問文がワードＡ，Ｂで概念検索する場合は，対象特許もワードＡ，Ｂで比較するから，このような問題は発生しない。

２．２　様々の概念検索

概念検索のすべてが上記の評価をしているわけではない。例えば②の観点の評価をしていないもの，つまりワードの重み付けの弱いものもある。また，データの蓄積時に類似度をあらかじめ計算しておくために上手な検索のために数ヶ月ごとにインデックスを再作成する必要のある方式もある。

さらに，最初に検索式で検索して数千件に減らしたものを対象に概念検索を行うものもあるが，これは，検索式作成の煩わしさはほとんど軽減されないことになり，特許情報活用における概念検索の役割りは限定的といえる。

また，日本語においてはワードの切り出しという基本的な問題もある。つまり，日本文は英文のように分かち書きされていないので，文章からワード（らしきもの）を切り出す必要がある。このやり方も，辞書による簡単だが新ワードへの対応が難しい切り出し方式とか，データベース全体の現状にしたがって利用時に切り出す（と同等の作業をする）方式など様々のものがある。

検索式方式においては，システムは異なっても and, or などの検索論理はほぼ同じであった。しかし，概念検索においてはシステムによって検索論理が大きく異なり，そのため検索結果は全く異なる場合がある。

文脈の意味を理解するような概念検索はまだない。このタイプの実用化は，特に日本語においては，相当先になると思われる。

３　概念検索の特徴

概念検索は，誰でも利用できる簡単な操作性と，短い所要時間が最大の特徴である。この特徴の認識は重要である。専門サーチャが何人もかかって（数）十時間かけて検討しつくした最高級の検索式による検索結果と，上手な質問文の作成法を知らぬまま，わずかな時間で行なった概念検索の結果を比較することがある。このような比較は少なくとも実務的利用の可能性の検討としては間違っているように思われる。概念検索は次々と改善されており，システムに適した上手な利用法を知っておく必要がある。

(1) 意味的に近い特許の検索が期待できる

例えば「液体で冷却するパソコン」を検索したい場合，従来の情報検索では「液体 and 冷却 and パソコン」というような検索式が使われて

1	2892	特開 2002-056236 インターネッ
2	2862	特開平 10-078992 自動競り方法
3	2850	特開 2001-060237 自動競り方法
4	2802	特開 2001-060238 競り方法およ
5	2693	特開 2001-319102 コンピュータ
6	2667	特開 2002-092387 インターネッ
7	2637	特開 2001-306879 医薬品相互情

(1)「インターネット競り取引で購入希望価格と，最大許容値を入力し自動的に競り取引をおこなう」で概念検索した上位７件。２，３番目の特許が該当。

1	3426	特開平 10-078992 自動競り方法
2	3411	特開 2001-060237 自動競り方法
3	3318	特開 2001-060238 競り方法およ
4	3258	特開平 07-302287 競りシステム
5	3121	特開 2002-183501 商品取引シス
6	3106	特開平 09-171531 自動電算卸売
7	3083	特開平 07-073251 自動電算卸売

(2)「購入希望価格と，最大許容値を入力し自動的に競り取引をおこなう」で概念検索した上位７件。いくつかのワードが欠けてもほとんど同じ結果を得ている。

図３　質問文の違いによる検索結果の比較

きた。これは対象とする明細書中（または要約中）に３つのワードを持っている特許を探せというものであり，指示内容が明確である。しかし，各ワードは重み付けされていないので，明細書中に一語でもあれば該当することになる。例えば，化学プラントなどの液体の冷却に関する特許で，ただ一回のみパソコンで信号を処理するというような記述のある，意味的には一致しないノイズ（不要な特許）も出てくることになる。

これに対し，概念検索では「液体で冷却するパソコン」という説明文そのものを質問条件とし，３つのワードでカバーされる概念を指定する。検索においては，切り出された３つのワードの様々な出現頻度の巧妙な点数処理がなされるので，必ずしも３つのワードを持っていない特許も他のワードの点数が高い場合は上位にランクされる。また，一回しか使われていないワードなどは点数が低くなり，例えば上記の化学プラントの例などは，点数が低くなり，ほとんど検索されないことになる。

(2) 同義語の追加が不要

概念検索では同義語の追加がほとんど不要である。全く不要というわけではなく，重要なワードは補った方が良いが，検索式の場合のように網羅することは不要である。これは，概念検索においては，**他のワードがある程度補ってくれることが期待できる**からである。

例として「インターネット競り取引で購入希望価格と，最大許容値を入力し自動的に競り取引をおこなう」という質問文を想定する。ここで，もし「インターネット」の「ッ」が大文字になって「インターネツト」になっている特許があったとすると，インターネットという質問ワードは無効であることになる。そうすると「競り取引で購入希望価格と，最大許容値を入力」だけで質問したことと同じになる。ところで，これらのワードで成り立つ概念を考えてみると，現在においてはインターネット競り取引以外の特許はほとんど存在しない。そのため，「インターネツト」となっている質問文でもほとんど同じ結果を得るのである。

図３は，一歩進めて「インターネット競り取引き」を入れた場合と，除いた場合の検索結果である。この検索結果もほとんど違いがない。このように，他のワードが同義語の役割を果たしてくれることが期待できる。

同義語において，もう一つの観点がある。「許容値」のワードを見ると，検索式に慣れた人は「許容価格」なども必要ではないのかと考える。しかし，これは不要である。「許容値」のようなワードでは，システムが「許容」と「値」に分解して処理してくれることが期待できる。このような意味でも同義語はあまり必要ではない。

さらに，同義語は不要というだけでなく邪魔になることがある。これは，特定のワードのみを同義語や上位概念，下位概念のワードを多数設定した場合，それらの重み付けを大きくするという誤った指示をする結果になる恐れがある。

※文章の質問が良い理由

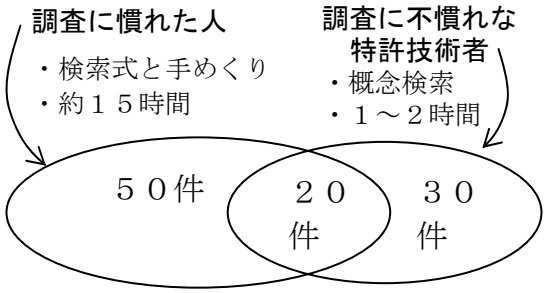

図４　検索式方式と概念検索の比較
・対象 48 テーマ件数は一定評価以上の合計。
・明らかに概念検索に不向きな質問文を使い回答０件だった 6 テーマを含む。

概念検索はワードに分解して処理するのであるから，本質的にはキーワードの入力でも問題はないように思われる。しかし，キーワードの入力では上記の例から推察されるように特定のワードを強調し過ぎる質問になるような誤った利用になる恐れがある。したがって，文法的に正確である必要はないが，文章で質問することを基本とし，必要なら一部をワードで補う質問が良い思われる。

(3) 概念検索は絞込みが不要

例えば，検索式方式ではある条件で 1,000 件も出るので，もう一つ and 条件で絞ると 10 件など少なくなりすぎるという問題がよくおきる。このために，and で絞るのは妥当か考え込むとか，また or で同義語を補うような悩ましい問題がある。ところが，概念検索では，質問文で与えられた条件にそって点数を付け，上位の数十件を出力してくれる。したがって，回答の多寡に悩む事はない。先頭数十件を見て，また次の質問文を決めればいいのである。

(4) 絞込み過ぎて 0 件になることはない

検索式方式では，絞り込みすぎて回答が 0 件に（近く）なることがあった。これを避けるために，同義語の検討や段階的な検索など悩ましいテクニックが必要であった。それに対して，概念検索では，質問文のワードに重み付けをした計算により点数の高い上位数十件を見るものである。したがって，質問文が絞りすぎた表現であっても，ほとんどのケースで問題にならない。例えば，質問文を「テレビの消費電力低減で，プログラムで制御し，深夜に低電力表示するもの」などとしたとき，全てを満足する特許がない場合は，いくつかの条件を欠いているものが最上位としてランクされる。

(5) ポイントを絞った検索に適している

概念検索は，対象とする技術概念を細かに絞り込んだ検索に適している。逆に，広がりのある技術の対象特許の全体を検索するような用途には適していない。全体を検索するような場合には，回答結果に優先順位をつけない，検索式による従来方式の検索が適している。

(6) 質問文の作成は技術者には易しい

質問文の作成は，エンドユーザには易しいがサーチャにとっては難しい。探したい情報を正確に表現するのは対象技術を熟知しているエンドユーザには易しく，そうでない専門サーチャには難しい。依頼書を読んで内容を理解するのに 1 時間前後かかり，さらにその技術分野で関連特許が多いのか少ないのかなど熟知するのにさらに 1 時間前後かかる。エンドユーザはその時間が皆無に近い。

(7) 操作は簡単

概念検索の利用法は 30～60 分程度の説明で十分理解できるほど簡単である。検索式方式は基本的なコマンドの習得に 1 日を費やし，さらに数日の習熟をしないと利用できなかった。

図４は，サーチャが時間をかけて検索式と手めくりで調査した結果と，調査に不慣れな特許技術者が概念検索をやった結果である。約十倍の時間の違いがあるのに，結果はあまり違わない。これは，概念検索が技術内容を熟知したエンドユーザに優位であることを示している。なお，上記の比較では特許技術者の方が調査対象を正しく把握し易かったことも考慮する必要があろう。しかし，このことがまさに，概念検索はエンドユーザの利用に適していることを表している。

(8) 概念検索はエンドユーザに適している

特許情報を利用する人は，様々な技術を対象に検索サービスを行なう専門サーチャと，研究者・技術者・特許技術者などエンドユーザと言

われる人に分けられる。サーチャはコマンドの利用法，検索式の作成法などを熟知しているが，依頼される技術の詳細を知らないので，調査依頼書を読み，予備検索をして技術内容を理解するのに１，２時間を要する。このため，検索に10時間前後かかることは一般的である。

これに対し，エンドユーザは，調査したい技術内容は熟知しているが，コマンドや検索式の作成法を知らない。このため，検索式方式の検索ではエンドユーザによる効果的な検索は困難であり，サーチャに依頼することが多かった。ところが，概念検索は検索式の作成が不要であり，コマンドらしいものはほとんどない。そこでエンドユーザ自ら検索することが可能になり，不慣れな初心者でも１時間程度である程度の回答を入手することができる。

参考文献；

1)特許情報検索の課題と概念検索システムの役割，六車正道，知財管理，Vol.51,No.12（2001.12）
2)特許情報検索における全文検索と概念検索の役割分担，市川伸治，特許庁／特技懇，No.223（2002.5）
3)企業における特許情報の活用，六車正道，特許庁／特技懇，No.223（2002.5）