概念検索はなぜ上手に検索できるのか？

ユーザのための概念検索の上手な利用法

六車技術士事務所・所長
六車　正道

PROFILE 約36年間、日立製作所において特許情報の活用企画と実務を担当し、2006年2月に独立して技術士事務所を開設し、特許情報の活用促進に関するコンサルタント業に従事。関係業務の講演や著作が多数。PatentCityの運営者。技術士（情報システム）。

http://ipbase.cool.ne.jp/mailmug.htmより　050-8012-2416

1 はじめに

概念検索を利用すると、簡単な質問文であるにもかかわらず目的とする特許がズバリ検索されて驚くことがある。

一方で、概念検索を上手に使いこなせず、使い物にならないと考えている人がいる。また、最初はすばらしいものだと思って使ったが、期待した結果が得られないために使い物にならないと考えるようになった人もいる。

現存する概念検索は、質問文やデータベース全体、また個々の特許文献でのワードの出現頻度を巧妙に使って類似度を計算している。しかし、それ以上の細部になると、原理がベクトル空間モデルとか確率モデルであると説明されたり、しかし実際にはさらに工夫して変形するために違いが分かり難くなったり、また、キーワードの長さや切り出し方の違いや類似度の計算法の微妙な違いなどの様々の工夫、さらに類似文書検索とか連想検索など表現の違いなどがあり、利用者レベルでの比較検討は容易ではない。しかしそれは、検索式のような単純なシステムでない知的な検索システムであるがゆえの代償ともいうべきものであり、今後ますます複雑になるであろう。とはいえ、基本的な事項はそれほど複雑ではなく、利用面からみると共通・類似のことがらも多い。（例えば、人の着る衣類でいえば、素材は異なり細部は違っても、着るものとして共通のことは多い。）

概念検索に関して、このJapioYearBookで2005年；基礎と使いこなし・システムごとの比較、2006年；知財業務での利用・研究開発でのアイデア発想支援的利用、2007年；概念検索についての誤解、として紹介してきた。本稿では、概念検索はなぜ長い質問文より短い質問文で上手に検索できるのかについて検討し、概念検索の使いこなし促進に役立つことを目指した。

2 概念検索はなぜ上手に検索できるのか？

2.1 概念検索の使い方

例えば「ハイブリッドカーで急発進時に電池の消耗を少なくする」ことに関する特許を探す場合、この文章を質問文として入力して概念検索する。そして上位数十件の抄録や明細書を目視チェックして最終的に自分の欲しい内容の特許をピックアップする。検索式利用でもそうであるが、概念検索でも、この目視チェックを避けることはできない。この例ではたとえば、過放電対策に配慮した特許をピックアップするなどである。

次に、目視チェックで知った知識で質問文を変えて、例えば「ハイブリッドカーの電池で急負荷時の過放電を防止する」などに変更して再検索を行う。さらに、観点を少し変えて「急激な負荷変動時のエンジンと電池のバランス」などの文章で概念検索を行う。このように4，5回繰り返すことでかなり再現率の高い結果を得ることができる。これが概念検索の典型的な利用法の1つである。

概念検索は「思いついたワードを入力」するだけで完璧な検索ができる魔法のシステムではない。

2.2 上手に検索できるとは？

上手な検索とは、正しい回答の多くを（つまり漏れが少なく）、回答リストに不要なものが少なく（つまりノイズが少なく）、低コスト（短時間、低料金）で取り出

す検索といえよう。このとき、漏れとノイズとコストは、一方を高く望めば他方が犠牲になるトレードオフの関係にある。望ましいわけではないが、短時間の検索では一定の漏れがあるのは当然である。（そうでなければ、長時間をかける調査・検索の必要がないことになる。）

概念検索では、その使いこなしの知識をもち、対象技術の知識のある人が行えば、1，2時間の短時間である程度の上手な検索を行うことが期待できる。しかし、多くの検索において言えることであるが、他の追加的な検索をすれば新たな正解を追加できるものであり、概念検索でも同じことが言える。

2.3　短文の質問文で上手に検索できる

概念検索ではどのような質問文で上手に検索できるのか検討する。表2-1は、特開2000-42356をターゲット特許としてそれに類似した特許を探した検索である。表の横軸に様々の質問文をとり、縦を検索対象としている。表中の件数は上位50番目までに出てきた内容的に一致した件数である。概念検索として比較的高性能である2つのシステムで比較している。

この概念検索では、検索の目的を「家庭から出る生ゴミを微生物で分解処理する際、排気を木材チップに通して脱臭する、木材チップの芳香と微生物で脱臭」とし、この短文を質問文として検索した場合、最も多く検索（Bシステムで6件、Eシステムで7件）できている。なぜ、このように短い質問文を使うと上手に検索できるのであろうか？

表2-2の左側は、上記の短文に出てくるワード分布、つまりワードとその出現回数である。なお、この表ではワードの点数の差を感覚的に把握できるように出現回数に比例して文字を大きくした。これらのワード分布から想像される技術は、検索目的に非常に近いことが理解できる。

概念検索では、これらのワードの or 検索がなされ、2回出てくる「脱臭、微生物、木材、チップ」と1回出てくる「家庭、生、ゴミ、分解、排気、芳香」が若干の点数の差で類似度が計算される。このため目的技術に近い特許が上位にランクされる結果になることが予想される。

表1　特開2000-42356（生ごみ処理の脱臭）の関連特許の概念検索

※上位50位までのヒット件数。元の特許を除く。

①Bシステムによるヒット件数

対象＼質問文	短文 54字	第1請求範囲 216字	要約 311字	明細書 10,800字
明細書	6件	0	0	0
請求+要約	6	0	0	0
請求範囲	4	0	0	0
要約	4	1	1	0

②Eシステムによるヒット件数

対象＼質問文	短文 54字	第1請求範囲 216字	要約 311字	明細書 10,800字
明細書	7	5	1	1
請求+要約	6	3	1	0
請求範囲	−	−	−	−
要約	−	−	−	−

短文；家庭から出る生ゴミを微生物で分解処理する際、排気を木材チップに通して脱臭する、木材チップの芳香と微生物で脱臭

第1請求範囲；タンク状の脱臭機本体と、上記脱臭機本体の内部を上方の脱臭室と下方の要処理ガス導入室とに仕切る仕切材であって、後記充填材の通過を許容せず、ガス及び水分の通過のみを許容する仕切材と、前記要処理ガス導入室に要処理ガスを導入する導入手段と、前記脱臭室からの処理済ガスを排気する排気手段と、前記要処理ガス…（以下、省略）

表2　短文と請求範囲に出てくるワードと出現回数

質問文中の出現回数

（2回）脱臭、微生物、木材、チップ
（1〃）家庭、生、ゴミ、分解、排気、芳香

質問文：短文

（6回）ガス、処理、脱臭
（5〃）導入
（3〃）仕切、充填
（2〃）通過、排気、許容
（1〃）タンク、木片、生、ゴミ、活性、汚泥

質問文；第1請求範囲

これに対し、同表の右側は第1請求範囲に出てくるワード分布である。このワード分布の技術、つまり、6回出てくる「ガス、処理、脱臭」や5回の「導入」が非常に高い点数となり、3回の「仕切、充填」などもやや高い点数となる技術は、検索目的よりもだいぶ異なるものがイメージされるのではないだろうか。

つまり、この実例のように、請求範囲では発明の重要部分を表すものでないワードが数多く使われるために、多くの場合、概念検索の質問文に適しないのである。

図2-1は、概念検索を行なう前に質問文から切り出されるワードを見ることのできるシステムで、ワード分布の認識のされ方を見たものである。「脱臭」や「木材チップ」などは質問文に2回出てくるので重み付けが高く、「生ゴミ」などは1回なので低いことが分かる。なお、このシステムでは、「木材チップ」と「木材」「チップ」や、「生ゴミ」と「ゴミ」などのようにダブって切り出されている。なお、このような切り出し方は最適の検索をするためのこのシステム特有の工夫によるものである。

なお、概念検索の最終的なスコアは、上記のような質問文での出現回数だけでなく、各ワードのデータベース全体での出現文献数の逆数や各特許での各ワードの出現回数などにより計算されるが、本稿では理解し易くするためにそれらを省略した。また、質問文からのワードの切出し方もシステムによって異なるが、ここでは同じワードにした。本稿の役割として精密な議論よりも、大筋を理解するものと位置付けて、このような説明にしている。

次に「要約」を質問文とした場合を検討する。日本公開特許の「要約」は発明者が書いたものであり、2，3行で示す【課題】と7，8行で書かれた【解決手段】から構成されている。表2-3左側はターゲット特許の要約であり、右側はその主なワードの分布である。このワード分布は請求範囲に近いものになっており、目的とだいぶ異なる技術になることが理解される。

次に、明細書全体を質問文にした場合について検討する。表2-4はターゲット特許の明細書に出てくる主要なワードの分布である。請求範囲や要約のワード分布と比較すると、「生」「ゴミ」、などの出現回数がやや多く、若干目的に近い内容に見える。しかし、上記の短文に比

提供；日立製作所／Shareresearch（シェアリサーチ）

特徴ターム(絞込設定)	重み付け	特徴ターム(絞込設定)	重み付け	特徴ターム(絞込設定)	重み付け
木材チップ	100	脱臭	100	木材	100
生物	100	チップ	100	微	100
生ゴミ	63	ゴミ	63	家庭	63
排気	63				

図1　質問文から切り出されたワードとその点数の例

表3　特開2000-42356の要約と主なワードの出現回数

【要約】【課題】 脱臭のために使用する材料が安全で容易に入手でき、運転操作が簡単で日常の管理に手間が掛からず、かつ安定した処理効果が得られ、設備費及び運転費のトータルコストが安価である生ゴミ処理機用脱臭機の提供。
【解決手段】タンク状の脱臭機本体1と、脱臭機本体1の内部を上方の脱臭室2と下方の要処理ガス導入室3とに仕切る網状の仕切り材4と、要処理ガス導入室3に要処理ガス5を導入する導入口11及び導入管12と、脱臭室2からの処理済ガス6を排気する排気口13及び排気管14と、要処理ガス導入室3の下部に構成した水抜き口15、水抜き管16及びその途中に挿入した水抜きバルブ17と、脱臭室2に充填材として充填される活性汚泥7を付着した多数の木片8とで構成する。

（7回）処理、脱臭
（6回）導入
（5回）ガス
（3回）排気、水抜
（2回）仕切、充填、運転
（1回）生、ゴミ、木片、活性、汚泥、バルブ、タンクなど

べるとはるかに請求範囲のワード分布に近い印象であり、やはり目的とだいぶ異なる技術になることが想像される。

以上のことから、多くの場合、長文においては利用者が着目した技術以外のワードが数多く使われることが多いため、質問文には適していないことが理解できる。

2.4 長文で上手に検索できる場合

ところが、質問文として明細書や請求範囲、要約などの長文でも上手に検索できることがある。なぜそのようなことが起きるのだろうか？

それは、長文のワード分布のうち上位のワードが目的とする技術内容のワード分布と類似している場合、といえる。先の例でいえば、長文の中で「脱臭、微生物、木材、チップ、家庭、生、ゴミ、分解、排気、芳香」が最も回数多く出てくる場合である。しかし、上記のワードも多いが他のワードがもっと多い質問文は不適当である、つまり、他のワードが多い特許を探すことになってしまう。

試みに、上記の短文と類似のワード分布になるような長文を作成して概念検索をおこなった。表2-5に示すのはその長文であるが、「脱臭、微生物、木材、チップ」を4回、「家庭、生、ゴミ、分解、排気、芳香」などを3回とし、多くのダミーのワードを1回ずつ、および平仮名や数字を配置した。

表2-6は、このやや長いサンプル文による概念検索の結果を、短文や要約による概念検索と比較したものである。これを見ると、サンプル長文の文字数は要約よりも多いが、ヒット件数は短文とほぼ同じである。

これらから推察されることは、質問が長文であっても目的とする技術関係のワードが「最も多い」ならば再現率は短文と同程度に高いということである。つまり、（明細書、請求範囲、要約などのような）長文が質問文として不適切ということは本質的なことではなく、長文では目的とする技術関係のワードが「最も多い状態にならない」恐れが多いので良くないという、確率的で実務上の問題といえる。

表4　特開2000-42356の明細書中の主なワードの出現回数

（162回）処理、（135回）脱臭
（120回）生、（107回）ガス
（55回）ゴミ、（47回）導入
（39回）排気、活性、（37回）汚泥、悪臭、（35回）木片、（30回）微生物
（29回）充填、（24回）仕切、（23回）水抜
（12回）家庭、（12回）通過、（11回）運転、（10回）分解
（6回）タンク、（5回）バルブ、（5回）芳香、（4回）許容
（0回）木材、（0回）チップ

表5　サンプルとして作成した長い質問文（文字数＝319字）

「まず1つめを4かいずつ、脱臭、微生物、木材、チップ、脱臭、微生物、木材、チップ、脱臭、微生物、木材、チップ、脱臭、微生物、木材、チップ、これからあとは2つめを3かいずつ、家庭、生、ゴミ、分解、排気、芳香、家庭、生、ゴミ、分解、排気、芳香、家庭、生、ゴミ、分解、排気、芳香、これから後は文字を長くするためのダミーとしてひとつずつせっていしたもの、冷却、加熱、混合、分割、知識、吸収、加工、薄型、剥離、合金、記憶、円盤、磁気、生体、反応、反射、容器、引出、認可、攪拌、時計、計測、病院、心拍、遠隔、接着、携帯、映像、文字、正確、練成、混合、分離、滅却、回路、光学繊維、カメラ、アルコール、メチル、ソフト、ハード、ストッパ、リミッター、プリンタ」

表6　サンプル長文のヒット件数
（対象；2004. 12/31以前の公開特許）
※上位50件目までの中で内容的に一致した件数

質問文 / システム	サンプル長文（319字）	短文（54字）	要約（311字）
Bシステム	7件	6	0
Eシステム	5	7	1

※検索対象は明細書。

3 概念検索を上手に使うための工夫

システムを上手に使いこなすには、そのシステムに合った使い方が必要である。テレビは携帯電話でも50インチの大型画面でも見ることができるが、目的や評価基準が異なることは容易に理解できる。しかし、概念検索と検索式の特徴を理解して、そのシステムに合った使い方をしているであろうか？

概念検索の細部はシステムによって異なるために、利用者レベルにおいて意見交換する基盤がない。さらに、操作法が極めて簡単であるためにそれが使い方の全てであると考えてしまう点にも問題がある。つまり、概念検索に対して自動的な検索を求めてしまい、使い方を理解して上手に利用することが大切であることについての意見が少ない。そこで、上手に使うための工夫を以下考察してみた。

3.1 思いつきの質問文での検索

思いつきの質問文で概念検索を利用する場合は、再現率は低くてもよいと割り切って使うとか、その後3，4回やり直すならば悪くない使い方といえる。

概念検索の宣伝文句を見ると「思いついたワードを入力するだけで検索できます」とか「特許番号指定で類似特許を芋づる式に検索できる」というようなものが多い。しかし、そのような利用法では、ある程度は検索できるが実務で満足して使うほどの再現率にならないことが多い。「このような使い方もできる」という程度に理解すべきである。ところが、これらの説明を聞いて、それで最良の概念検索ができると早合点し、使ってみて再現率の低さに驚き、概念検索はまだ時期早尚と考える人が多い。概念検索は「思いついたワードを入力」するだけで完璧な検索ができる魔法のシステムではない。

他人に調査を依頼するときは言葉を選んで慎重に説明し、調査員からも質問があり必要な事項に収斂していく。ところが、現在のコンピュータ検索システムでは、（検索式でも概念検索でも）コンピュータが利用者に質問を出すことはない。したがって、利用者が検索結果を見てやり直すことが必要である。概念検索は検索式に比べてやり直す手間が少ないので、数回のやり直しをしてもそれほどの負担にはならない。

3.2 40文字程度以上の質問文が良い

質問文として、短すぎるものは不適当である。質問文が「安価な液晶表示装置」では、該当する特許が多すぎて、利用者が必要とする観点に絞りきれない。ましてや「モータ」など1ワードの質問では単にそのワードを数多く持つ特許が最優先に表示されることであり、概念検索に適した使い方をしていないと考えるべきであろう。

一定の技術内容を指定するには、現在の1,000万件前後のデータベースの現状では、40文字程度以上の質問文にするのが妥当なようである。

3.3 80文字程度以下の質問文が良い

長すぎる質問文では、その中のどこにコンピュータが着目するか分らない。つまり、どのようなワードの出現回数が多いのか予想できない。よって、長過ぎる質問文は使わない方が良い。

また、特許番号を指定して類似特許を検索するための概念検索は、指定した特許の要約や請求範囲や全文を質問文として概念検索を行うものである。したがって、最適の短文の質問にくらべると、多くの場合、高い再現率は期待できない。

とはいえ、長文や特許番号指定での概念検索は簡単に使えるので、高い再現率を求めない場合や、その後何度か概念検索を行うのであれば、それなりに役に立つ。

3.4 同義語で質問文を分ける

質問文に同義語や類義語を入れ過ぎると、それらのワードに偏った結果になり、目的の技術に絞り込まれないことが生じる。「入れ過ぎ」とは、短文の場合、2，3個以上といえる。

概念検索では他のワードが同義語の役割をしてくれることが期待できるので、同義語を設定する必要性は低い。検索式では同義語を増やすことは、ノイズを増やさぬ限り、必須のことであり、増やし過ぎても問題はない。しかし概念検索では、同義語を入れるとそれらのワードの配分が高くなって偏りが大きくなり、再現率を悪くする

ことが多い。

とはいえ、産業日本語（技術用明晰日本語）のような標準化された用語で特許文書が書かれるようになれば、概念検索の再現率はいっそう高まるといえる。

概念検索で同義語を入れたい場合の対策・・・そのワードを使った別の質問文で検索すると、トータルとして再現率を向上させることができる。

3.5　具体的なワードによる質問文がよい

質問文に使うワードとしてはできるだけ具体的なワードがよい。上位概念のワードでは、類似の他の技術との区別ができず、上手に概念検索できないことが多い。具体的なワードにした場合、類義語を補うことが必要に思われるであろうが、前項で説明したようにそれほど必要ではない。

3.6　絞り込む検索に使う

概念検索は、質問文に対する類似度を計算し優先順にリストアップされる。したがって、優先順を付けずに網羅的にリストアップする検索には適していない。これを間違って、例えばA社のインクジェットプリンタの動向分析のために全件をピックアップするような業務に使うと役に立たないことになる。

概念検索は、出願前の検索とか無効資料を探す検索などピンポイントに絞り込んだ検索に適している。ただし、短時間の概念検索だけで高い再現率を期待するのは限界がある。

3.7　50番目程度までは見る

概念検索の回答は、コンピュータの評価により内容的に当たりの可能性の高い順にリストアップされるが、実際に使ってみると、かなりの割合で1位でないところに最も近いものが出ている。また、ちょっとした文章の相違で順番が大きく動くこともある。これは、1番目という場所はただ1ヵ所しかなく、例えば2〜10位までの場所は9ヵ所もある、その一方で、例えば1〜10位の点数の違いは少なく、ちょっとした文章、つまりワードの違いで順位が入れ替わる。ということであれば、2〜10位までの方が9倍もの場所があり、1位よりは2〜10位に最も近いものが来ることも多いと考えてはどうだろうか。

なお、目視チェックで最終的に必要なものに絞ることは重要である。例えば、いくつかの部品を重ねた順序の違いを質問文で指示してもそれは検索できないので目視チェックで絞るしかない。これは検索式の場合と同じである。

そのようなこともあり、多くの場合上位50番目あたりまではチェックが必要であり、絞りが弱い質問文ではもっと多くまで見ることが必要なこともある。何件まで見るのが妥当かということは利用者が決めることである。

4　おわりに

筆者は今年、概念検索の指導を数多くのユーザ企業におこなった。開発部門を対象に1社で2桁に及ぶ説明会を行なうとか時間をかけて実例研究をおこなうなど、それ以前の単なる講演形式よりも1歩進んだものを要望された。単純明快な検索式に対して、ワードに重み付けを行なうなど知的な側面を持つがゆえに複雑で見方によっては分かり難い概念検索はこのようにして徐々に浸透していくものであろう。

データベース検索は、寡黙な調査員に検索を依頼することに似ている。概念検索（寡黙な調査員）から依頼者に質問が出されることはない。依頼者は質問事項を正しく伝える工夫をし、結果を見て何度かやり直せば、比較的簡単に良い結果が得られよう。概念検索は何度かやり直すことが容易に行なえる。

特許情報を概念検索で利用することで、関連技術における先人の創意工夫を簡単に知り、学ぶことができる。概念検索はそのような利用に適しており、知財関連の業務だけでなく研究開発担当者のアイデア発想支援ツールとしても役立つ。このような利用は特許制度の本来の目的のひとつであり、今後、研究開発の様々な面で利用が進むことを期待したい。