論　　説

概念検索システムの有用性の研究と提言

知的財産情報検索委員会
第　2　小　委　員　会*

抄　録　2000年に初めて日本特許の概念検索システムが提供され，2004年までに更に数システムで概念検索がサービスされるようになり，これまで以上に概念検索システムを利用できる可能性が高くなった。本報告では現在提供されている５つの概念検索システムについて仕組みや機能の概要と事例による検証結果について報告し，現状の概念検索システムの性能について紹介する。

目　次

1．はじめに

1．1　目　　的

日本特許を検索対象とする概念検索システムは2000年にNRIサイバーパテントデスクから初めて提供された。従来型のブール演算による検索システムとは異なり，検索条件を自然語による文章で与える画期的なサービスであり，エンドユーザのための検索ツールとして脚光を浴びた。情報処理技術の進展に伴い近年次々と概念検索機能を追加した検索システムが登場してきた。これまで概念検索システムはブラックボックスで何をしているか分からないと言われ，調査専門家からは敬遠されがちであった。当小委員会ではこのブラックボックスといわれた概念検索の仕組みの部分を少しでも明らかにし，ユーザの理解を深めたいと考える。なお，今回の研究ではシステムの優位性ではなく，概念検索システムの仕組みの解明と現状の性能について全般的に評価することを目的としている。

1．2　用語の定義

各システムでは機能や特徴について共通表現を用いていない場合があるため，ここでは便宜的に表現を統一した。概念検索とは，検索条件を自然語による文章で与え，システムがその検索条件文を分析し，適合する案件を検索結果として抽出し，検索結果を検索条件に類似している順にランキング表示する機能を指す。システムによっては類似検索とも呼んでいる。

また，検索条件文を種文書，検索条件文や蓄積文書から抽出される特徴的な語句を特徴語（または特徴文字列）と呼ぶ。

* 2004年度　The Second Subcommittee, Intellectual Property Information Search Committee

2．概念検索システムの仕組みと特徴

今回の調査対象は主要な５システム（NRIサイバーパテントデスク，PATOLIS-IV，ATMS，RIPWAY，Shareresearch）とした。概念検索システムの基本的な動きとしては，まず種文書・蓄積文書とも形態素解析を行い，特徴文字列を抽出する。その後は大きく２つに分かれ，抽出した特徴文字列を使用して，各蓄積文書に固有のベクトル値を事前に計算して保持しておくものと特徴文字列の頻度データや重みを保持するものとがある。前者は種文書が与えられると同じようにベクトル値を算出し，種文書のベクトル値に近い順に蓄積文書をランキングする。後者は頻度データに基づいた重みを基に類似度を計算し，検索結果をランキングする。概念検索システムの仕組みと特徴の概要については表１に示す通りである。サービス開始時期や改良が最近のものほどいろいろな機能が追加されている。例えば，概念検索の際に検索条件を文章で指定するだけでなく，IPCなど一部の書誌事項を同時に指定することができるものや，ユーザの適否をフィードバックし再度ランキング処理できるものなどがある。

3．事例による概念検索システムの評価

3．1　無効資料調査のための特許テストコレクション

当小委員会では2001年度～2003年度にかけて国立情報学研究所（NII）と共同で２つの特許検索システム評価用テストコレクション（「技術動向調査のための特許テストコレクション」「無効資料調査のための特許テストコレクション」）を作成している。今回，事例による概念検索システムの有効性の検証を行うに当たり，

表１　概念検索システムの方式・機能

	ATMS（ジー・サーチ）	NRIサイバーパテントデスク（NRIサイバーパテント）	PATOLIS-IV（パトリス）	RIPWAY（リコー）	Shareresearch（日立）
データ登録時の処理	抽出語を生成し頻度データ，重みを計算	各文献に固有のベクトル値[*1]（要約，クレームで別）を計算	特徴語を抽出し，その頻度データを保持	形態素解析[*2]・係り受けを解析しインデックス作成	抽出語を生成し，頻度データを保持
索引方式	n-gram[*3]と単語インデックスのハイブリッドシステム	非公表	極大単語索引[*4]	デュアル索引[*5]	インクリメンタルn-gram[*6]
概念検索の流れ	種文書から抽出語を生成 ↓ OR条件で要約およびクレームを対象に全文検索 ↓ ランキング処理	種文書から単語を抽出 ↓ 種文書のベクトル計算 ↓ ランキング処理	種文書から特徴語を抽出 ↓ OR条件による全文検索 ↓ ランキング処理	種文書から形態素解析で単語を抽出 ↓ TF-IDFによる重みの高い上位15語を抽出 ↓ ランキング処理	種文書から特徴n-gramを抽出 ↓ 種文書のベクトル計算 ↓ 抽出語を含む案件を対象にランキング処理
ランキング処理	TF- IDF	種文書のベクトル値との内積で類似度を計算	辞書の頻度データおよび重みを基に類似度を計算	確立モデル	TF-IDFに基づいたベクトル計算
概念検索用辞書	頻度と重みを計算したDB	ベクトル辞書 年１回更新	出現頻度辞書 リアルタイム更新	適合文書に共出現する関連語を拡張語として検索語群に追加	単語辞書は持たない。但し，頻度辞書あり
同義語辞書	なし ユーザによる登録機能有	なし	なし	なし	同義語・異表記展開
語の重み付け変更	不可	不可	可	不可	可（マイナスも可）
語の追加/削除	可/切り出し語から選択	不可	可	可/ユーザフィードバックによる関連語拡張	可
望ましいと思われる種文書長	—	100文字以上	技術内容を表した文書100文字以上	請求項３つ程度	２～３行の短文
検索対象期間の限定	公報発行日	不可	不可	出願日，公開日，登録日，優先日	出願日，公報発行日
日付以外の検索条件との組合	出願人，IPC	IPCメインクラス	全検索項目（但し，新たな検索集合は類似度データなし）	IPC，FI，Fターム，出願人，発明者，代理人	ワード，IPC，FI，出願人，権利者，発明者，代理人
概念検索の履歴閲覧・保存	検索条件を保存。検索結果の通常検索への流用も可	不可	不可	同一セッション内で検索結果，検索条件を保存（最大５つ）	同一セッション内で検索結果，検索条件を保存
概念検索対象	要約＋請求の範囲	要約，請求の範囲	要約，請求の範囲，全文	要約+請求の範囲，全文	要約，請求の範囲，全文
特徴点	語句抽出において，品詞選択や不要語指定が可能	唯一，空間ベクトル法を採用	パトリスフリーキーワード辞書を使用	ユーザの適否判断をフィードバックし再ランキング可能	同一セッション内での概念検索の履歴保存

*1　特徴語の出現頻度を基に文章をベクトルで表現したもの
*2　文章の文節，品詞，接続関係などを解析する技術
*3　n文字単位で文章から文字を切り出しインデックス化する方式
*4　単語索引をベースに切り出す単語に冗長性を持たせる方式
*5　n-gramと単語索引を用途により使い分ける方式
*6　切り出すｎ文字を可変にする方式

当小委員会の成果物である無効資料調査を前提とした検索課題と適合特許リストからなる「無効資料調査のための特許テストコレクション」を使用した。これは，現在の概念検索システムは，網羅的に関連特許を探すというよりは，短時間で関連特許をいくつか探すという目的で利用されているケースが多いと考え，このような利用実態に近い事例としてこのテーマを選定した。また，検索課題数が34個と多く，適合特許が事前に判明していることから，限られた時間内でシステムの評価を行うのに適している。更に課題特許の技術分野を筆頭IPCセクション[1)]でみた場合，A～Hまでの8セクション中，Dセクションを除く7セクションがカバーされているという点でも適当であると判断した。34課題とこれら課題に対する適合特許344件の技術分野（筆頭IPCセクション）の分布を図1に示す。

テストコレクションでは，課題特許を無効化する案件（適合特許）の対象範囲を1993年～1997年の公開特許（公表，再公表は含まず）に限定している。このため，これ以外の期間及び公表/再公表は適合特許抽出対象範囲ではないことから事例検証での評価対象から外している。

適合特許はA判定とB判定の2段階に分かれており，A判定は単独で潰せる案件，B判定は構成要素の一部が欠けているため，他の情報と組み合わせることで無効化可能となる案件である。また，適合特許の明細書内における適合箇所（パッセージ）とその組合せについても抽出している。

テストコレクションの詳細については，『知財管理』Vol.53 No.5 2003を参照のこと。

3. 2　事例検証に利用した課題と検索条件

事例検証に用いた検索条件は特定の公開特許番号の特定請求項とし，当該特許の出願日若しくは優先日以前に公開されていることとした。34課題の特許番号，基準日（出願日若しくは優先日）と対象請求項は表2に示す。ほとんどの課題に於いて請求項1が対象請求項となっている。

3. 3　事例による検証方法

事例検証の対象システムとして，NRIサイバーパテントデスク，PATOLIS-IV，ATMS（但し，改良後の検索エンジンを使用），RIPWAY，

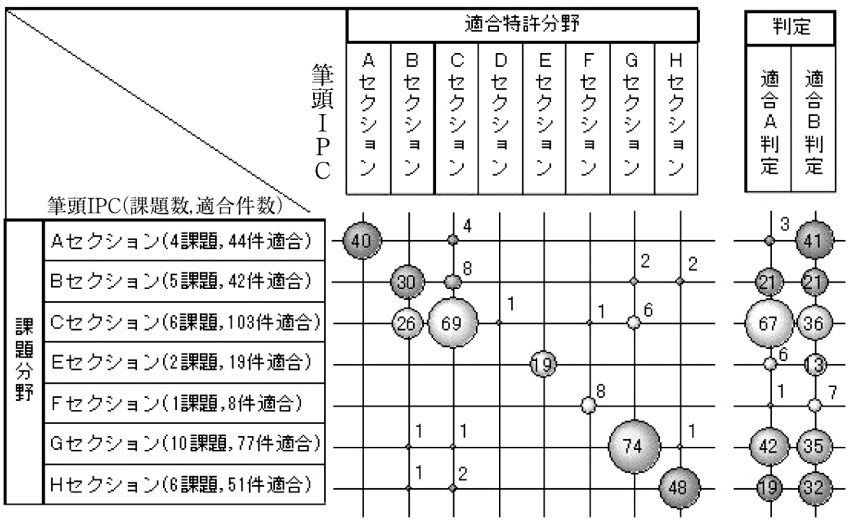

図1　IPCセクションから見た課題と適合特許の関連

表２　課題一覧

公開特許番号	発明の名称	基準日	対象請求項
特開平07－099887	食用油用添加剤	1993/9/30	1
特開平07－164776	感熱孔版印刷原紙用フイルム	1993/12/16	1
特開平07－258008	水田用除草剤組成物	1994/3/25	1
特開平07－281442	平版印刷版用現像液およびそれを用いた製版方法	1994/4/13	1
特開平08－048860	熱可塑性ポリエステル樹脂組成物	1994/8/9	1
特開平08－074325	建物ユニットの連結部材	1994/9/1	1
特開平08－076374	ポジ型フォトレジスト組成物及びパターン形成方法	1994/8/31	1
特開平08－092055	美白化粧料	1994/9/22	1
特開平09－051209	誘電体基板および配線基板	1995/8/8	1
特開平09－164527	成形機の樹脂材料供給装置	1995/12/14	1
特開平09－195289	補強土壁面ブロックおよび補強土壁面構造	1996/1/23	1
特開平09－256119	無方向性電磁鋼板およびその製造方法	1996/3/21	1
特開平09－265470	電子ブック装置	1996/3/28	1
特開平09－318650	センサ装置及びその製造方法	1996/5/27	1
特開平09－328646	印刷インキ組成物	1996/4/8	1，3
特開平10－048123	分光分析装置	1996/8/6	1
特開平10－067695	不飽和アルデヒドおよび不飽和カルボン酸の製造法	1996/8/26	1
特開平10－090986	静電潜像現像装置	1996/9/12	1
特開平10－138831	車両モニタシステム	1996/8/6	9
特開平10－141062	ディーゼルエンジン	1996/11/7	1
特開平10－200720	ファクシミリ装置	1997/1/13	1
特開平10－323893	スチレン系樹脂組成物を使用したインフレーションフイルム又は２軸延伸フイルム	1997/5/26	1
特開平10－329409	記録材	1997/5/30	1
特開平11－002834	液晶表示装置	1997/6/12	1
特開平11－066711	記録媒体識別装置	1997/8/22	1
特開平11－084093	Ｘ線用多層膜光学素子	1997/9/10	1
特開平11－176552	端子と導体の接続方法	1997/12/11	1
特開2000－044535	スルホニウム塩及びこれを用いたフォトレジスト用添加剤	1998/7/16	1
特開2000－067869	非水電解液二次電池	1998/8/26	1
特開2001－031508	ヒョウダニの忌避剤	1999/7/16	1
特開2001－111675	携帯電話装置	1999/10/8	1
特開2001－236994	多孔性ポリマー電解質の製造方法及びそれを用いた非水電解質電池	2000/2/24	1
特開2002－162775	静電潜像現像用トナー	2000/11/29	1
特開2002－167651	マルテンサイト系ステンレス鋼及びその製造方法	2000/11/16	1

SGPAT（Shareresearchと検索エンジンが同一）の５システムで行った。検索は各システムにおけるデフォルトの条件下で行った。また，今回は検索条件が請求項であるため，検索対象領域を請求項と明細書全文（以下，全文と略す）の２種類とした。しかし，システムによっては，検索対象領域として抄録と請求項が一緒になっているものもある。

また，期間限定できない２システムについては，上位3,000件を抽出し，その中から今回の調査対象の期間・公報種別の案件についてのみ再抽出し，手作業で再ランキングした。このため，基本的には上位500位までを評価対象としているが，システム・課題によっては500位までのデータが抽出できないものもあった。このように，各システムの検証データは全く同じ条件によるものにはなっていないが，可能な限り条件を揃えるよう努力した。

評価に使用したデータは，請求項（一部要約を含む）を対象にした検索結果５システム，全文を対象にした検索結果３システムの計８種類で行っている。評価は検索結果の上位500位にテストコレクションの適合特許344件が抽出された割合（再現率）を検証している。

4．評　　価

(1) 全般的に請求項対象より全文対象の方が結果は良い。

請求項を対象にした結果と全文を対象にした結果を比較すると，上位100位までの適合特許の再現率は，請求項対象では平均35%，全文対象では平均41%となり，全文対象の方が全般的に結果は良かった（図２）。これは，適合特許における適合箇所が請求項以外でのケースも多くあり，通常，それらは全文対象でなければ拾われてこないためと考えられる。多くのシステムでは，まず種文書から切り出された複数の特徴語のOR条件により１次抽出がなされる。この１次抽出に掛かるためには，特徴語が１回でも多く出現する割合が高い方が有利であり，文書長が長い全文対象の方がその可能性が高くなるといえる。

また，個別システムでみると上位50位以内にランクされる適合特許の再現率は，請求項対象では22～37%，全文対象では31～42%となり，両者とも10%以上の差があることが分かった。しかし，いずれのシステムでも上位100位までを見ることはかなり効率的であるといえる。

(2) 適合特許の請求項数の違いによる影響

適合特許の請求項数の違いによる検索結果への影響度を検討した。適合特許の請求項数を１～２個，３～５個，６個以上の３区分（以下，請求項数区分と称す）に分けて比較した(図３)。

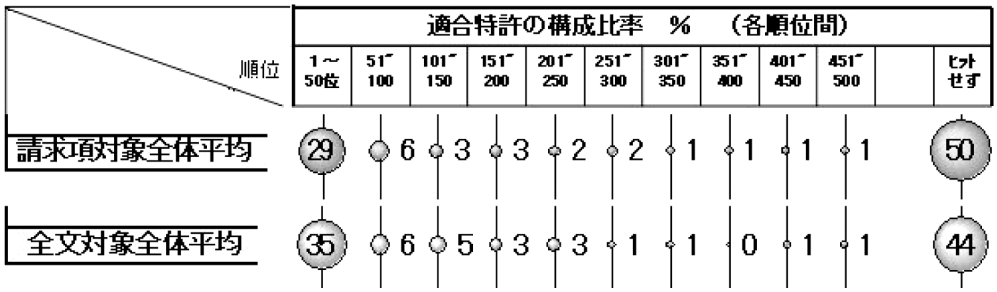

図２　上位500位までの適合特許の抽出状況

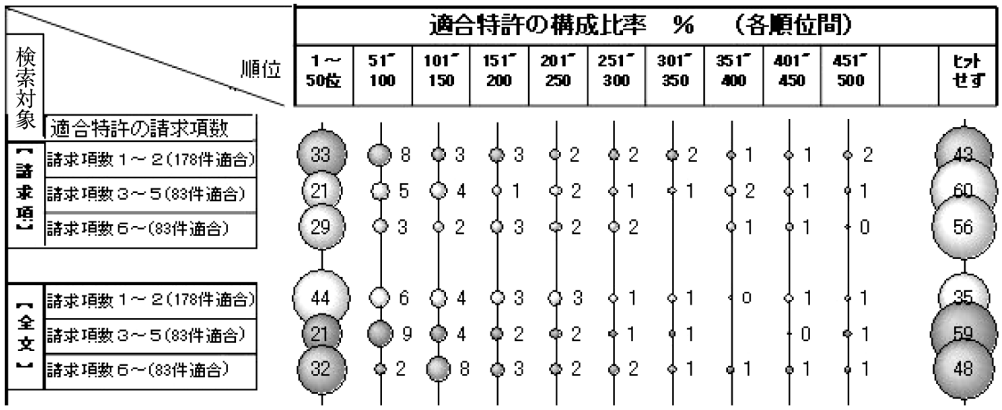

図３　適合特許の請求項数の違いによる適合特許抽出状況

小委員会では，適合特許の請求項が少ない方が概念は明確になりやすいため，適合特許の請求項が少ない方が結果は良いだろうと予測していたが，必ずしも請求項数が少ない方が良い結果になるとは限らなかった。請求項数区分別にみた上位500位における適合特許の再現率は，全システムの結果の全体平均で請求項数１～２個→６個以上→３～５個の順に良かった。また，適合特許の請求項数による影響について技術分野別に検証した結果，技術分野により傾向が異なることが分かった（図４）。例えば，Ａセクションでは請求項が少ないほうが結果は良いが，Ｆセクションでは逆の傾向が出た。

(3) 課題特許の技術分野の違いによる影響

課題特許の技術分野別に適合特許の抽出状況を比較した（図５）。請求項対象，全文対象ともＡセクションは比較的良く，Ｆ，Ｈセクションは悪い。但し，各種の要因により，必ずしもその通りにはならない場合もある。例えば，技術分野による影響について，課題特許の技術分野別に適合特許の請求項数区分別の抽出状況を検証してみたが，各請求項数区分にて，結果が良い技術分野は必ずしも一致しないことが，分かった（図６）。

(4) 条件文の語と一致性が高い適合特許は上位にランク

適合特許のうち上位にランクされるものとランク外のものを比較しランキング結果の根拠を検討した。概念検索結果とテストコレクションのパッセージを突き合せた結果，条件文と適合特許の請求項（あるいは明細書）中の語の一致性が高いものは上位にランクされ，逆に異表記のように意味としては同じでも語が一致しなければランク外となる。例えば，課題特許ではディスプレイ，携帯電話装置，と記載されているが，ランク外の適合特許では，それぞれ，表示手段，通信端末，と記載されているなどである。異表記以外では，特徴語の出現頻度の影響が大きいと思われるものがあった。

また，パッセージの該当箇所（請求項かそれ以外か）と調査対象範囲（請求項か全文か）が単純に影響するケースと，影響しないケース（例えば，パッセージの該当箇所が実施例であるのに，全文対象結果ではランク外で，請求項

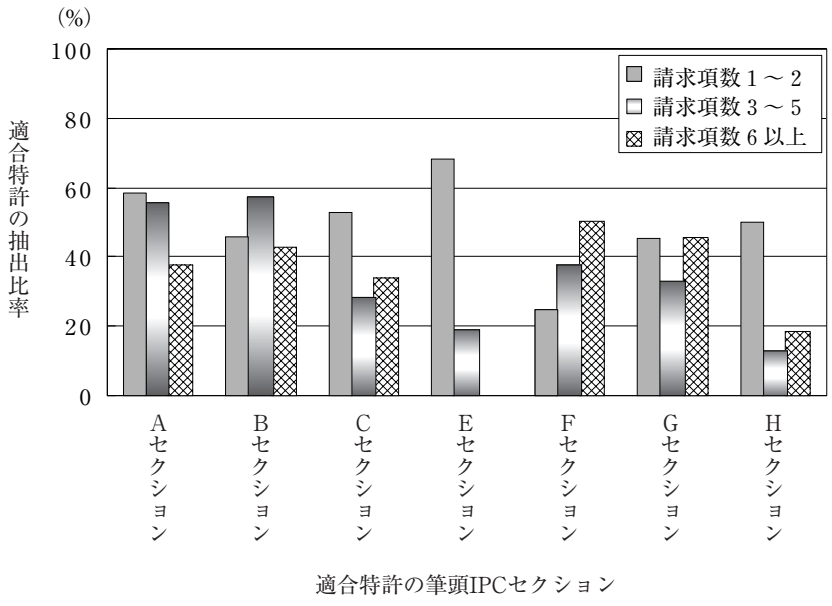

図４　適合特許の請求項数の違いによる分野別抽出状況

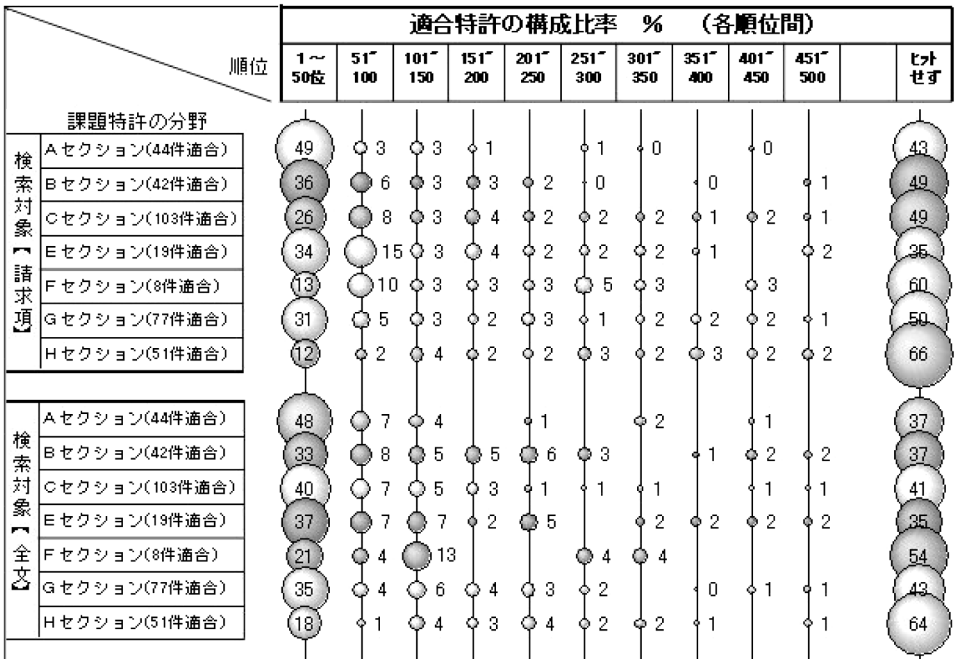

図５　課題特許の分野別適合特許抽出状況

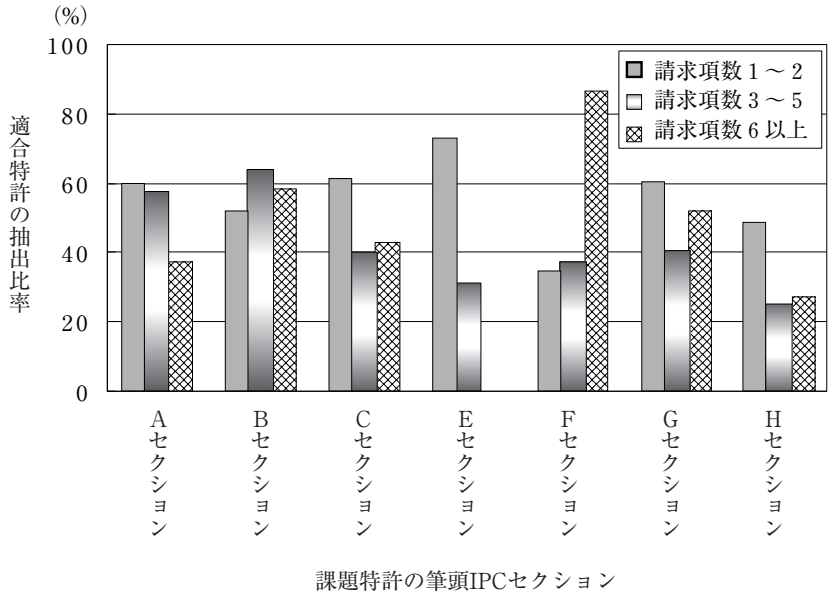

図６　課題技術分野（筆頭IPCセクション）における適合特許の請求項数区分別抽出状況

対象結果ではランク内)があることが分かった。

(5) A判定の分野別抽出状況

今回の事例検証では無効資料調査のための特許テストコレクションを利用していることから，上位50位までにA判定適合特許がランクされる課題数に注目し，課題技術分野（筆頭IPCセクション）別に抽出状況を比較した（図７）。その結果，システムにより７セクション全てでランクするものとAセクション，Fセクションが抽出できないものとに分かれ，AB判定両方で検証した場合よりも，システム間の相対評価

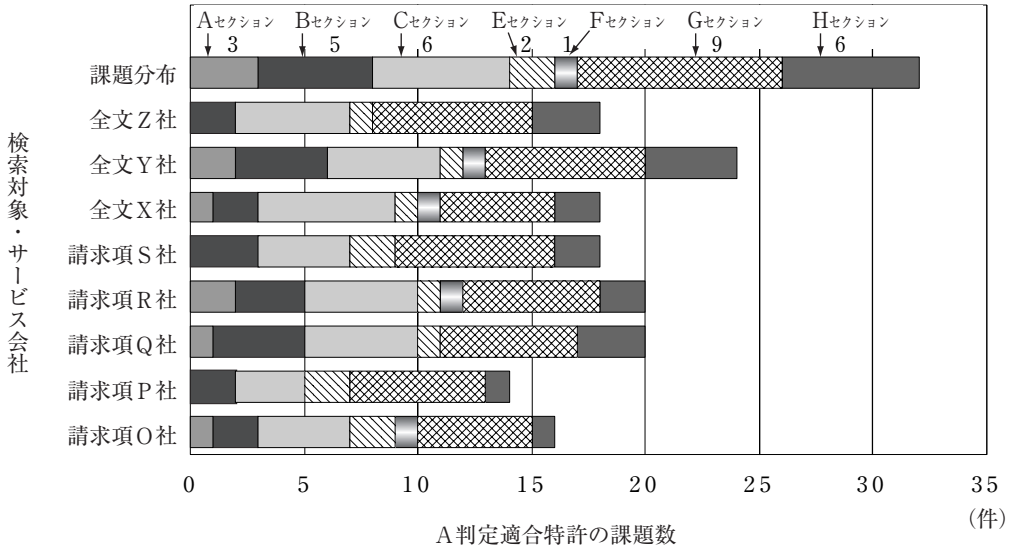

図7　課題技術分野（筆頭IPCセクション）別A判定適合特許の抽出状況（上位50位）

で向上するシステムや，全文対象よりも請求項対象の方が相対評価で向上するシステムがあることが分かった。なお，34課題のうちA判定適合特許を有する課題数は32個となっている。

5．まとめ

以上の評価結果から，上位50位若しくは100位までの結果を見ることはかなり効率的である。また，条件文で使用する語は結果に及ぼす影響度が大きいため，注意が必要である。目的や観点を変えることでシステムの評価は異なる。

今回の事例検証を通して，サービス業者への要望は次の通りである。①今回のような無効資料調査では日付による限定が不可欠であるため，日付も絞込み条件として追加していただきたい。②同義語・異表記･類義語への対応をお願いしたい。③今回の事例では技術分野間の差異が大きかったので，技術分野間での精度差が最小になるようにチューニングをお願いしたい。

今回，概念検索システムの現状サービスの性能評価に当たっては，特許テストコレクションの存在は大きかった。同じ課題数を独自に作成するには適合特許の判断を行う膨大な時間が必要であり，もし同じ課題数を独自に作成していたら到底一年間の活動内では実現できなかったことである。

一般にテストコレクションは情報検索システム分野の研究開発進展のために作成されており，日本におけるテストコレクションの作成は，国立情報学研究所が評価型ワークショップの形態で新聞情報や論文情報などを対象に行っている。しかし，特許情報については，多くのサービスが提供されていることや特許情報活用の活発化を考慮すると，システム開発のためだけのデータではなく，我々ユーザがシステムの性能評価や条件文の工夫等の試行錯誤をする際のデータとして非常に有効に活用できるものである。このように特許テストコレクションの存在の有意性を考えると，特許庁を始めとして国の機関が中心となり，より有効な特許テストコレクションを作成し，広くそのデータを利用できるようにしていただきたい。そうすることで，特許情報システムの飛躍的な性能向上が期待できるとともに，特許情報の活用が促進されると考える。

6．おわりに

概念検索システムの仕組みの解明や事例検証に際してご協力をいただいたシステム提供各社各位に深く感謝する。

なお，本研究テーマに携わった2004年度知的財産情報検索委員会第２小委員会委員は以下の通りである。堀越節子小委員長（日本電気特許技術情報センター），小川幸文（JFEテクノリサーチ），川本敦子（東芝），岸井晶三（積水化学工業），倉田昇（松下電器産業），柴裕昭（京セラミタ）。

また，無効資料調査のための特許テストコレクションの作成に関わった2003年度知的財産情報検索委員会第２小委員会委員は以下の通りである。仲村栄基委員長代理（積水化学工業），堀越節子小委員長（日本電気特許技術情報センター），松谷貴己小委員長補佐（日本化薬），池内覚（コニカミノルタ），奥田浩司（ポリプラスチックス），川口厚（ソニーテクノリサーチ），倉田昇（松下電器産業），坂木守（キヤノン），柴裕昭（京セラミタ），馬場健次（堀場製作所），三輪保（第一製薬），安田吉宏（新日本製鐵）。

注　記

1）IPCセクションとは，IPC（国際特許分類）の分類体系における最上位概念を表す階層で，全技術分野がＡ～Ｈの８セクションで大別されている。詳細はIPC分類表参照。

参考文献

(1) 岸田和明．情報検索の理論と技術，図書館・情報学シリーズ３，頸草書房，1998.

(2) 知的財産情報検索委員会第２小委員会，特許検索システム評価用テストコレクションの作成と評価結果，知財管理，Vol.53，No.５，2003.

（原稿受領日　2005年６月６日）