論　説

過去分公開公報テキストデータの評価方法の検討

知的財産情報検索委員会
第 3 小 委 員 会*

抄　録　過去分（CD－ROM公報以前）公開公報のテキストデータ化が，特許情報提供事業者から種々行われ始めている。これら過去分のテキストは，紙公報をスキャンニング後，OCR処理したものである。事業者によっては，目視確認をしている場合もあるとのことである。しかし，具体的な精度を正式に公表しているものはない。このOCR精度が，テキスト検索精度と直接関係するため，検索者にとって非常に気になるところである。利用する立場からこれら過去分公開公報のテキストデータの信頼性，有用性を検証するための手法を紹介すると共に，これら過去分テキストの検索での使用方法について提案する。

目　次

1．はじめに

平成５年以降の公開公報は，CD－ROM（または，DVD）で提供されている。つまり，電子データで提供されていることになる。そこで，これらの公報については，全文テキスト検索は，様々なデータベースで使用できている。それより以前の公報の場合，紙媒体での提供のため，検索手段としては，IPCやＦタームなどの分類検索，またはパトリス抄録，パトリスフリーキーワード検索しかなかった。

しかし，ここ数年，公報の全文テキスト検索ができる期間を遡らせるデータベースが登場してきた。これらは，データベース事業者が自社でOCR処理をしているか，他社がOCR処理したデータを購入（使用許諾）しているものと思われる。

本検討の主目的は，過去分公開公報テキストのOCR精度の評価手法を探ることである。どのデータベースの精度がよく，どのデータベースの精度が悪いといったことを示すものではない。ある特定の少ない検索タームにおいてのヒット率を確認する方法を示しただけである。調査目的や検索タームの種類・結果の見やすさなど，利用者の趣向による部分も多いことから，最終評価はデータベース利用者各位のご判断にお任せしたい。

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

2．検討内容

2. 1　評価手法

OCR精度を評価する方法として，以下の方法を考えた。

ある検索タームについて，対照用として，電子化登録（公告）公報を対象として，あるデータベースで検索を行いヒットした公開番号リストを記録する。また，目的の過去分公開公報を対象として，各データベースで検索し，公開番号リストを記録する。これらをマッチングさせ，電子化登録公報でヒットした案件において，過去分公開公報がどれだけヒットしたかを数値化（以下，ヒット率という）することによって，検証できるのではないかと考えた。

実際には，公開公報が発行された後で，拒絶対応等により補正が行われ，登録公報になる場合も多い。しかしながら，認められる補正として，公開公報に存在しなかったタームが，登録公報に含まれることは希であるとの前提からのものである。このことを確かめるために，比較として，平成５年の公開公報と電子化登録公報でのマッチング結果も確認することにした。この場合は，どちらも電子化公報であるので，純粋に審査段階等の補正による影響が確認できる。

データベースによるが，「一度のCSVダウンロード件数が1000件以内である」，「ダウンロード費用が発生する」などとの制約等から，12個の検索タームでの評価となった。

また，全文検索の範囲，文字の統一化など各データベース検索エンジンの違い（特徴）により，その影響を受けた結果となってしまうことは避けられない。それでも，個々のヒット率はかなり正確と考えるが，n数が少ないことから「木を見て森を見ず」のごとく，各データベースの正確な全体像把握には不十分な可能性があることを最初にお断りする。

なお，本研究のための検索は，2004年６月～2005年１月に行ったものである。

2. 2　評価対象

調査対象データベースは，下記のものとした。

DocuPat（富士ゼロックス），HYPATWeb（発明通信社），JP－NET（日本パテントデータサービス），NRI（NRIサイバーパテント），PATOLIS（パトリス）

検索期間・対象は，昭和58年～平成４年の公開特許公報とした。公表や再公表特許については，未収録データベースがあるため除いた。また，一部のデータベースでは，「昭和58年までの提供がない」，「検索に費用が掛かる」，などの理由から検索できていない公開年がある。

検索タームおよび文字種分類は，下記とした。

普通漢字：架設，介挿，埋込

複雑漢字：緻密，撥水，輻射

カタカナ：ピラン，ブチロ，ワニス

英字：ACI，HCL，MGO

多種ある文字種のうち，検索で使用する頻度の高そうな文字形式から選択した。さらに，漢字に関しては，OCRを考えた時に，精度の差があるかもしれないので，普通漢字と複雑漢字とに分けた[1)]。

各文字種分類におけるタームは，ヒット件数が適当なものを試行錯誤して選択した。ヒット件数が少ないと，検証精度が低下する。逆に，ヒット件数が多いと，時間と費用が掛かる。

また，英字は，同一視させているデータベースとそうでないデータベースがある。この影響を避けるため，全ての大文字・小文字の組合せで検索した。例えば，ACI＋ACi＋AcI＋Aci＋aCI＋aCi＋acI＋aciとした（DocuPatの英字検索は標準では単語単位の検索となるので，それらの前後一致検索で行った）。

2. 3　対象期間遡及のために

当初，対照とする電子化登録公報を平成５年以降（実質的に平成６年以降の公告公報または登録公報）のデータからターム検索した。

その後，平成４年の公開公報から徐々に遡って，検索期間の拡大を試みたが，平成６年以降の公告・登録公報では古い公開公報のデータは極端に少なくなるため，精度が十分に得られなくなることが分かった。そこで，遡及版特許公告公報CD－ROM[2)]データを検索対象とできないか検討した。

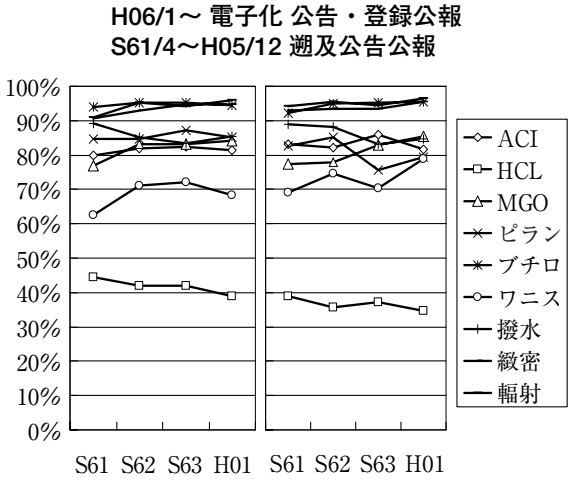

図１　電子化公報と遡及公告公報の比較

対照用として使用したい遡及版特許公告公報と電子化公報のデータで，比較できるほどの件数がある昭和61年から平成１年の間で比較したところ，ほぼ同等な結果が得られた。（図１；縦軸はヒット率，横軸は公開年）

遡及版特許公告公報のテキストは製本用キャラクタデータを元に作成されたので信頼性が高いと考えられるが，図１の確認結果で，両方のデータによる差があまりないことからもそれがうかがえる。従って，遡及版特許公告公報も電子化登録公報とほぼ同等に対照用として使用できるものと判断し，昭和58年の公開公報までの遡及データについて検証することとした。

2. 4　検索ロジックの詳細

今回のヒット率算出式は以下のようになる。

集合１：電子化特許登録（公告）公報および遡及版特許公告公報にて調査

全文（或いは，請求の範囲）＝検索ターム

and公開日＝S580101～H051231

and（（登録番号 >＝2500000）or

（公告日>＝19860401））

集合２：公開特許公報にて調査

全文＝検索ターム

and公開日＝S580101～H051231

集合３：集合１ and**集合２**

⇒実際の処理はPATOLIS以外では，公開系と登録系のand演算ができないので，公報番号をキーにマッチング処理を行った。

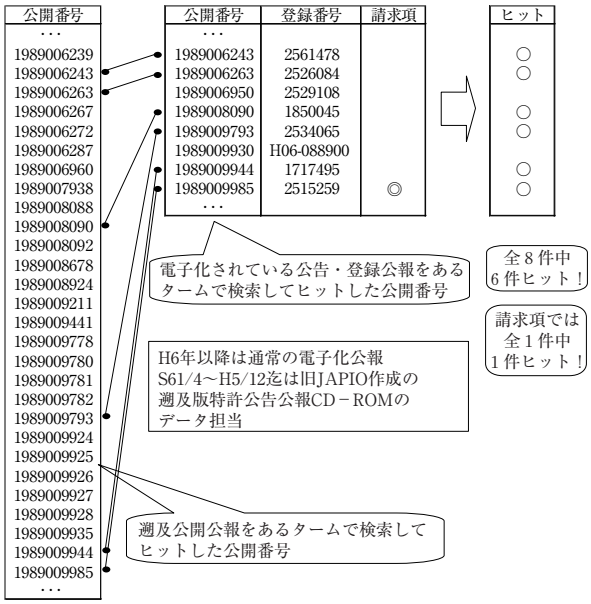

図２　マッチング処理の例

従って，**ヒット率＝集合３**（両方にヒット）／**集合２**（登録公報ヒット）となる。

2. 5　各ターム結果

各タームにおける公開年別ヒット率の推移を順に示す。５データベースを，任意にＡ～Ｅと

して表した。全文ヒット率は，全文でのヒット率であり，請求範囲ヒット率とは，対照用公報の請求の範囲に該当検索タームが含まれているものにおいて，ヒットした割合を示している。検索により公開公報に本タームが含まれていたかは確認していない。

また，検索タームの出現頻度は考慮していない。つまり，１公報で同じタームが10回出現していた場合，そのうちの10回全部でヒットしてもヒットと判定されるし，そのうちの１回しかヒットしていなくても，ヒットとなる。逆に，誤変換によりヒットしないと判定された場合には，10回出現していた公報上の10ターム全ての誤変換でも，１回しか出現しない公報の誤変換も，同じヒットしない公報１件となる。

個々のタームの結果例として「ACI」に対する各データベースの結果（図３～４）を示す。

同様に，他のタームでもヒット率を求め，それぞれ文字種分類毎で平均した結果(図５～12)を示す。

普通漢字（図５～６）の場合には，請求範囲のヒット率が全文よりも低くなっている。公開公報本文記載のワードが，登録公報では，請求の範囲へクレームアップされているためと思われる。これは，電子データ化されている平成５年の公開公報と登録公報の比較結果が同様であったことからも推定できる。この場合のように，平成５年データが100％に近くないなら，このタームは公開と登録との間の補正の影響であるので，OCR精度を確認するためには，平成５年データを基準に，それ以前の年のデータを読み直す必要がある。

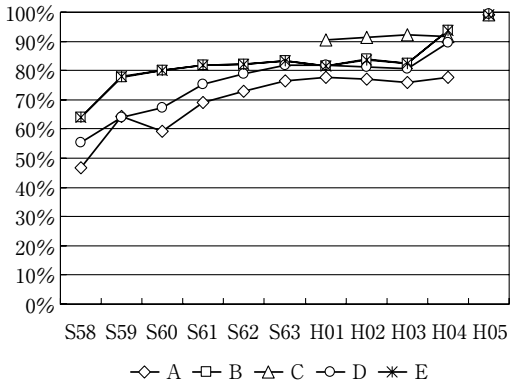

図３　ACI全文ヒット率

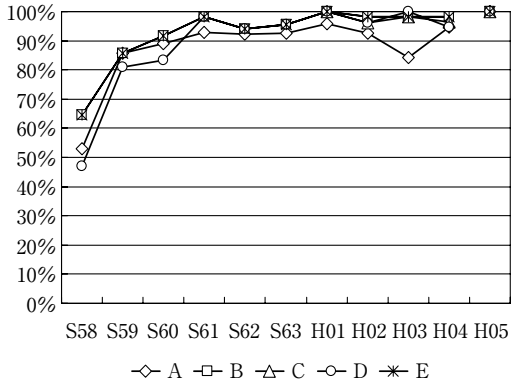

図４　ACI請求範囲ヒット率

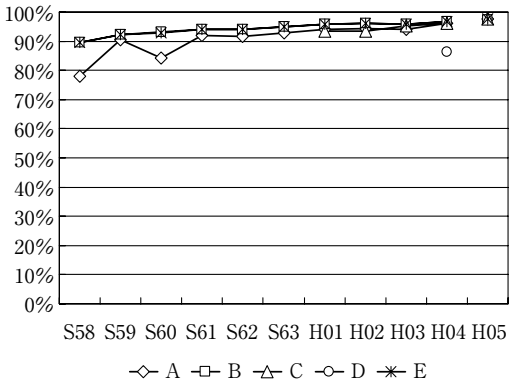

図５　普通漢字平均全文ヒット率

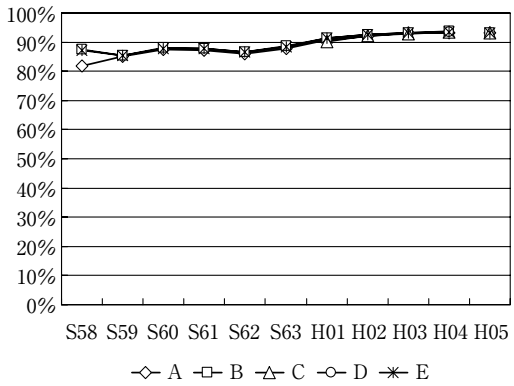

図６　普通漢字平均請求範囲ヒット率

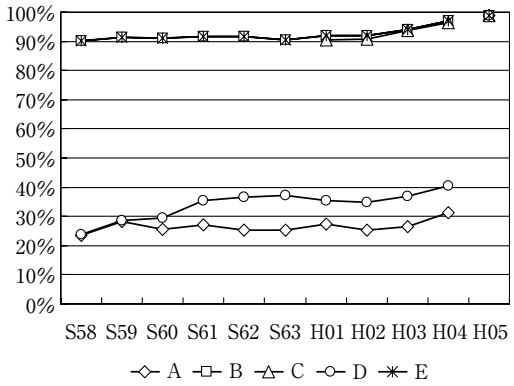

図7　複雑漢字平均全文ヒット率

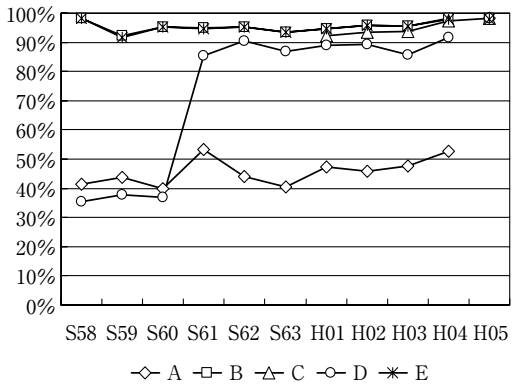

図8　複雑漢字平均請求範囲ヒット率

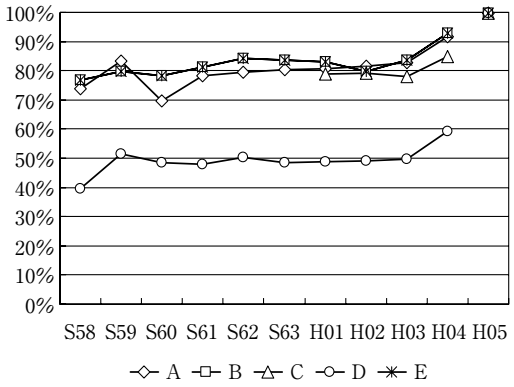

図9　カタカナ平均全文ヒット率

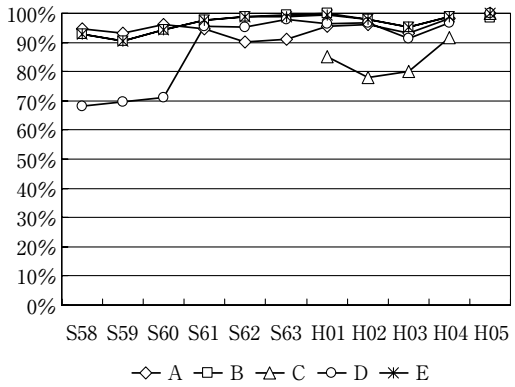

図10　カタカナ平均請求範囲ヒット率

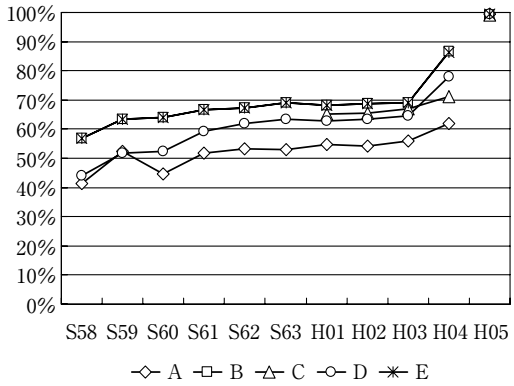

図11　英字平均全文ヒット率

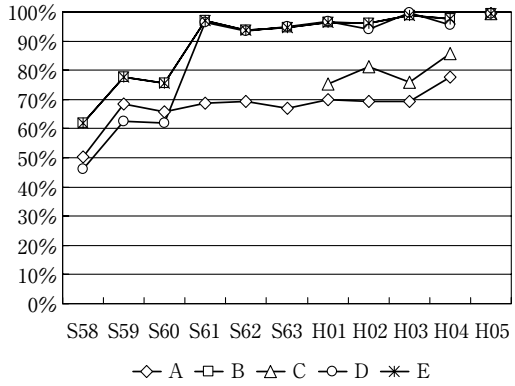

図12　英字平均請求範囲ヒット率

2. 6　誤変換の確認

HClの一認識例として特開平01－042486を取り上げる。この公報は１つのデータベースのみでヒットしたものである。

（CHCl3）の記載が公報に14ヶ所あるが，(CHCQ3)，(CHCl23)，(CHCQ3)，(CHCQ3)，(CHC113)，(CHCQ3)，(CHCQ3)，(CHCQ3)，(CHC123)，(CHCQ3)，(CHCQ3)，(CHCQ3)，(CHCQ3) および（CHCQ3）と認識し，１ヶ

所のみ正しく認識している。筆記体の「l」で記載されているため，多くは「Q」と誤認識されている。(図13)

３α－カルボン酸　ｐ－ニトロベンジ
８１mgを得た。
赤外吸収スペクトル（ＣＨＣℓ３）
$\nu_{C=O}$ （cm^{-1}）＝１７７５、１７４

図13　誤認識された紙公報例

次に同じHClでの偶然の誤ヒット例を紹介する。特開平01－068751には検索タームは４ヶ所存在するが，その部分は全て誤変換している。(下線部が正しくはHCl)

反応物を５％IIcNで中和した。
72時間攪拌し，５％11Clで中和した。
反応混合物を５％IIC／で中和した。
反応混合物を５％IIC／で中和した。

ところが，偶然違う場所で，誤変換したものが偶然HClとなって検索でヒットしている。

アミノ基例えばNH2，NHCH3又は
↓
アミノ基例えばNllz，NHCl13又は

このような状況を図示すると図14のようになると思われる。

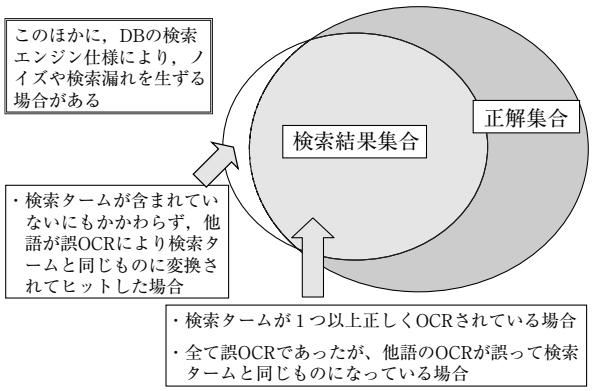

図14　検索ターム ヒット状況図

誤認識しやすい文字の一例として，下記のものが挙げられる。

「へ」「へ」「ぺ」「ぺ」「べ」「べ」←ひらがな，カタカナの半濁音，濁音
「り」「リ」←ひらがな，カタカナ
「I」「l」「1」←アイ，エル，いち
「ニ」「二」←カタカナ，数字
「エ」「工」←カタカナ，漢字
「ロ」「口」←カタカナ，漢字
「ー」「一」「-」「-」「一」←長音，ハイフンの全角および半角，漢数字
「.」「,」「.」「,」←全角及び半角のカンマ，ピリオド
「0」「O」←ゼロ，オー
１Ｌｌｌｉ（）{}[]¦;:!←縦に直線っぽい文字
ＯｏａＣｃＤｅ＠○←丸っぽい文字

また，誤変換の起こりやすい原因として，半角英数（記号・カタカナ）２文字分を，全角１文字と認識していることもある。半角と全角が認められる日本語特有のOCRの困難さを表している。さらに，公報中の表などで文字が小さなものや縦向きで記載されているものも誤変換の要因と思われる。当然ながら，手書き文字の公開公報や，一部公表公報の小さな文字では認識率は大幅に低下することが予想される。

2. 7　まとめと考察

これらに基づき，各データベースにおける文字種分類毎の平成１年～４年の結果をまとめると，以下のようになる。公開公報平成５年データが100％になるように，補正した値とした。

全文	A	B	C	D	E
普通漢字	98%	98%	98%	90%	98%
複雑漢字	30%	95%	95%	40%	95%
カタカナ	85%	85%	80%	50%	85%
英字	55%	75%	70%	70%	75%

請求範囲	A	B	C	D	E
普通漢字	99%	99%	99%	99%	99%
複雑漢字	50%	97%	95%	90%	97%
カタカナ	95%	97%	80%	95%	97%
英字	70%	97%	80%	97%	97%

本手法は，簡易的検証方法であるが，想像以上に各データベースにおいて，差が見られた。

古い年代ほどヒット率が低下傾向にあるが，その程度は比較的少なかった。特定データベースである年度以降では，請求の範囲に検索タームが存在する場合は，存在しない場合に比べ精度が高い。このことは，請求の範囲部分を目視確認により修正しているデータベースにおいて，精度向上の効果があることが分かった。ただ，一般に請求の範囲記載のタームは，本文中でも繰り返し使用されることが多く，その出現頻度により，高めのヒット率を誘発している可能性もある。

また，同一データベースでも，普通漢字・複雑漢字・カタカナ・英字という文字種での差が見られた。普通漢字は，ヒット率が相当高そうであり，今回の５種類のデータベースにおいて，十分に検索に使用できるであろうことが分かった。しかしながら，一見簡単そうなカタカナ，英字は，相対的にヒット率が低く，検索する際には考慮する必要がありそうである。複雑漢字は，データベース間の差が一番顕著であった。同一データベース内で，カタカナ，英字，複雑漢字で比較すると，カタカナを得意とするＡデータベース，英字を得意とするＤデータベース，複雑漢字を得意とするＢ，Ｃ，およびＥデータベースといった傾向がある。

一般に，OCRの認識率とは，OCRした文字数のうち正しく認識できた文字の割合を示す数字であり，全て正しく認識できた用紙の割合ではない。従って認識率99％でOCR処理された100文字の請求項があった場合，そのうち１文字が誤変換ということになる。常に同じ処理が行われたとして，その１文字を含んだ検索タームを使用した場合は，ヒットしない。つまり，認識率99％といっても，いろんな認識率の違いがある文字種の平均であることになる。つまり，OCR認識率99％と聞いて，該当公報100件中，99件は検索により摘出でき，誤変換により検索漏れが生じる公報が１件だけであると考えるのは，大きな間違いであることになる。

2. 8　結果に基づく使用方法

ヒット率80％であれば，OCR精度はある程度高そうに思える。ある程度の検索漏れを覚悟した検索なら使用できそうである。しかし，実際の検索を行う場合には，検索タームを組み合わせるため，ヒット率はさらにそれらの掛け合わせとなる。３つの検索タームの場合，80％×80％×80％＝51％である。従って，テキスト検索のみに頼るのではなく，IPC，FI，Ｆタームなどの分類検索を利用しながら，最低限のテキスト検索を組み合わせることが望ましい。また，近傍検索の利用も考慮すればよいと思われる。

テキスト検索をする場合，同義語のor検索をすることは知られている。OCRテキスト検索を使用する場合には，さらに検索したいタームの誤変換しやすそうな文字列とのor検索をする方法も有効である。例えば，「濾過」は，公報中には「ロ過」（カタカナのろ）と記載されている場合がある。これを，あるデータベースで，「口過」（漢字のくち）で検索することによって，「濾過」，「ろ過」および「ロ過」でヒットしなかった公報が見つかった。

本研究において，OCR処理された過去分公開公報のテキスト検索に適したタームと適さないタームが存在することが分かった。例えば，普通漢字は適したタームであったが，英字や複雑漢字は適さないタームであった。検索したいタームにおいて，過去分と平成５年以降分との

ヒット件数の比較や，電子化登録公報とのマッチングによるヒット率を確認するなどの予備検索をすることで，検索したいタームが過去分テキスト検索において，適切なものであるか判断されることをお勧めしたい。

簡単な方法として，例えば，OCR処理されている平成４年の公開公報と，電子データ化されている平成５年の公開公報におけるヒット件数だけで比較してもある程度のことが分かる。あるデータベースでこの２年間のヒット件数の差が大きい場合には，当該データベースと検索タームの組み合わせは検索に不向きであることを示唆している。

データベースの違いまで確認したい時には，ノイズ発生のないデータベース同士では，単純にヒット件数のみで相対的な優劣を比較できそうである。一般に，検索タームとしては，意味ある文字列を使用することになり，その文字列の一部の文字で誤変換が起こる可能性があることになる。一部の文字が誤変換された時に，その誤変換を含む文字列が意味あるものとなることは珍しい。すなわち，誤変換OCR文字列は意味のない文字列となって，検索タームとなることは少ない。従って，誤変換によって，ヒット件数が減ることはあっても，偶然ヒットすることは少ないはずである。このことから，ヒット件数の多い方が，OCR精度が高いと予想される。但し，この場合正解の件数は不明のため，ヒット率までは分からない。

さらに，最大限に漏れがない検索が要求されるならば，可能な限り多くのデータベースで検索を行い，いずれかでヒットしたものを利用することができる。例えば，「HCL」はヒット率が低く，昭和60年の公開公報では10～40％である。しかし，４種類のデータベースのいずれかとした場合は，60％となる。また，５種類のデータベースの使用が可能な平成１年の場合，各々は20～50％であるが，いずれか１種類でもヒットするのは70％となる。単独で一番高いヒット率であるデータベースよりも，ヒット率が20％向上したことになる。（図15）

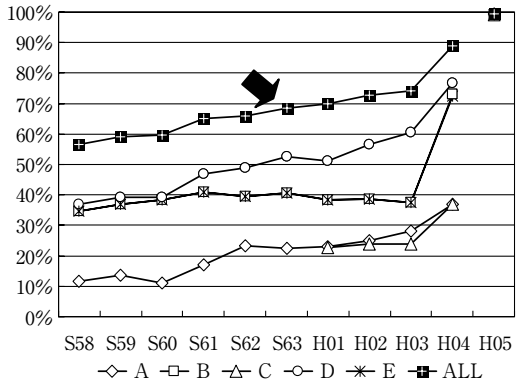

図15　HCL結果で５DBのいずれかでのヒット

3．ま と め

今回の調査手法には，以下のような問題点も含んでいる。

(1) のちに電子化登録（公告）公報が発行されたものに限定される。→公開公報のうち，３割程度しか評価対象とならない。

(2) データの入手の手間やコストが大で，マッチング処理も煩わしい。→このため，多数の検索タームを我々ユーザーが分析することが困難である。

(3) ノイズ発生が多いデータベースではヒット率が高く算出される恐れがある。→誤変換されていても，ノイズによりヒットする場合がある。

しかしながら，簡易的には過去分公開公報のOCRテキスト精度を確認でき，実際に使用する場合にはどの程度使用できるものであるかは，分かった。また，検索者としても，検索に使用する際に注意すれば，ある程度使用できることも把握できた。

データベース提供事業者側にも，本研究データに対する追加検証などを含め，自らの評価をしていただき，さらなる精度・使い勝手の向上

をお願いしたい。また，OCR精度の向上のみでなく，異表記や用語の統一化処理・シソーラス辞書の拡充など，ユーザーにとって便利な機能を付与することを，検討していただきたい。

本紹介では，省略させていただいたデータもあり，さらに他の手法で検証した研究成果もある。これらは，後日発行のCD－ROMに収録予定である。

4．おわりに

本研究は，2004年度知的財産情報検索委員会第３小委員会の下記のメンバーによるものである。

西井貞男（小委員長：チッソ），新井隆史（三菱樹脂），今津均（ノリタケカンパニーリミテド），酒巻由美子（シチズン時計），永山真実子（石川島播磨重工業），および中出良治（委員長，アドバイザー：エムテック）によるものである。

注　記

1）　画数の多い旧漢字などを複雑漢字と定義した。

2）　昭和61年４月～平成５年12月までに発行された公告特許のテキストデータを，収録したもの。

（原稿受領日　2005年５月31日）