英語・原語によるハイブリット検索：
PatBase,QPAT(Orbit.com),Discover による英語・原語ハイブリット検索の検討

○田畑文也[1)]，石田政司[2)]，水町保宏[3)]
富士フイルム株式会社[1)]，株式会社　神戸製鋼所[)2]，株式会社　アイピックス[)3]

〒421-0396　静岡県榛原郡吉田町川尻 4000　富士フイルム株式会社[1)]
Tel: 0548-34-5401　FAX: 0548-32-8286
E-mail: fumiya_tabata@fujifilm.co.jp

Hybrid Patent Search by English and Original Language:
Hybrid Patent Search by English and Original Language on PatBase,QPAT(Orbit.com) and Discover.

TABATA Fumiya[1)], ISHIDA Seiji [2)] , MIZUMACHI Yasuhiro [2)]

Fujifilm Corporation [1)], Kobe Steel, Ltd. [2)], IPICS Corporation [3)]
Fujifilm Corporation [1)] ,4000, Kawashiri, Yoshidacho, Hibara-gun, Shizuoka , Japan
Phone: +81-548-34-5401　Fax: +81-548-32-8286
E-mail: fumiya_tabata@fujifilm.co.jp

【発表概要】

中国特許検索など非ラテン語圏を対象国とする特許調査において、網羅性が必要な調査では、商用データベース（DB）の英語検索に加え別途原語 DB の補完が必要になる場合が多い。 このように異なる DB を用いた場合、検索結果の統合を含め、工数が膨大になる。

昨年来、商用 DB である PatBase、Orbit.com (QPAT)、Discover が英語検索機能に加え、原語による検索機能をリリースしたことにより、一つの DB で、英語・原語の検索ワードを混合したハイブリッド検索が可能となった。 これらの DB の実際のデータ収録状況の検証を行い、更にこのハイブリッド検索機能を活用することにより、原語 DB での補完と同様の効果が得られるかを検証したので以下に報告する。

【キーワード】

ハイブリッド検索，多言語検索,マルチ言語検索，原語検索，中国語検索

1. はじめに

近年、新興国での特許出願が急激に増大し、特に中国の2010年度特許出願件数は、ついに日本を超えて世界第二位になった。さらには、中国は実用新案の出願も多く、特許と実用新案の 2010 年度出願総数は約 80 万件となっている。更に出願件数に占める内国人出願率が約 3/4 である現状を考えると、この内国人出願もうまく検索で捉える必要がある。(図.1)

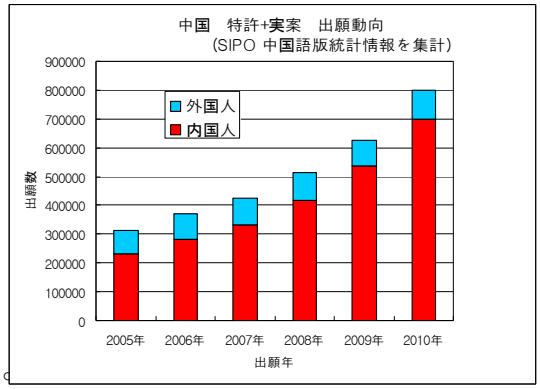

図 1.　中国　特許+実案総数　出願数動向

一方、中国の内国人出願が外国出願されるケースは少ないため、商用データベース(以下 DB)の対応特許でこれらの中国内国人出願を把握することが難しい場合も多い。

今までは、中国特許について、網羅性の高い検索が必要な場合、商用 DB の英語検索に加え、データ収録、タイムラグ、機械翻訳精度の問題から、別途 CNIPR 中国語 DB などでの原語検索による補完が必要であった。[1)]

しかし、CNIPR 中国語 DB では、DB のレスポンス速度、安定性に加え、検索結果のダウンロードに問題があり、有料 ID であっても、効率的に大量の件数を取り出すのが難しいという問題がある。

また、複数の DB の結果を特許番号形式を変換した後、マージさせる作業等も必要で、確認作業を含め、工数も膨大になる。

昨年から、いくつかの商用 DB において、英語検索だけでなく、原語のキーワード、出願人等も検索可能な機能がリリースされた。さらに、その他いくつかの DB でも、マルチ言語での検索機能付与リリースの予定と聞くが、2011 年 7 月末時点で、日本に上市されている商用 DB のうち、英語に加え、中国語などの非ラテン語の全文検索機能に対応しているのは以下の3種類である。

① PatBase (RWS 社)
② Orbit.com、QPAT (Questel 社)
③ Discover (CPA Global 社)

上記商用 DB の内、中国特許のデータについて詳細にデータ収録、機能、精度などを調べたので報告する。

尚、本稿では、一つの DB での英語+原語(中国語など)の検索をハイブリッド検索と呼ぶこととする。ハイブリッド検索の例として、Orbit.com の検索例を図.2 に示す。

Orbit.com ハイブリッド検索画面例

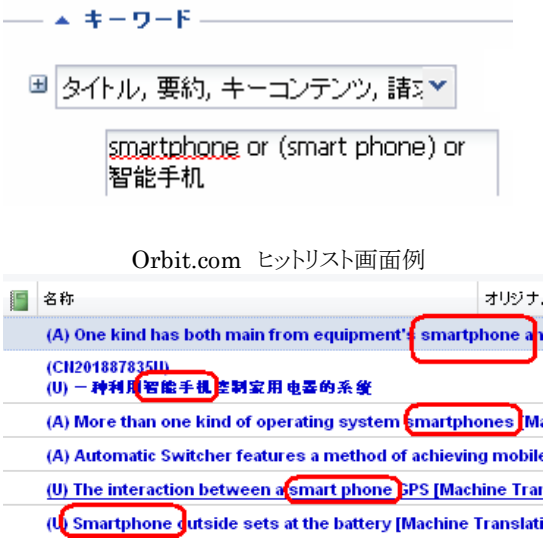

図.2　Orbit.com ハイブリッド検索画面例

2. 検証内容

ハイブリッド検索の例として、中国公開特許を対象に、DB 収録について英語、中国語の両方のデータについて、収録を確認し、実際に検索できるか等を確認した。なお、Orbit.com(QPAT)については、FAMPAT（ファミリーベース）を選択し、評価した。

また、DB 毎に機能が異なり、全く同じ検索条件で比較できないところも多く、一部は代替手段（目視で抄録の有無などをカウント等した項目）も用いている。

従って、同じデータ項目でも、評価方法が一部異なり、検証精度が良くない部分もあるので、留意されたい。

2-1 中国公開特許の集計方法

中国公開特許の限定は以下の検索式例で実施。1990 年以降発行分について、発行年毎、2010 年 6 月以降発行分については、月毎の収録を確認した。また、収録率の基準件数として、CNIPR 中国語 DB の件数を用いた。尚、PatBase、Orbit.com(QPAT)については、ファミリー数、Discover については公報ベースの件数で集計した。

PatBase: (CCD=CNA201010)
Orbit.com:((CNA L 2010)/PN)
Discover:
PD:range(2010-01-01,2010-12-31)
PC:("CN") KD:("A")

2-2 データベースの英語抄録収録率

PatBase、Orbit.com については、収録確認用コマンド(AB=YES 等)を用いたが、Discover については、英文抄録の有無の確認コマンドないため、伊藤らが INFOPRO2009 で報告した汎用語を用いた方法で確認した。[2)]

Discover: AB:("com*" or "con*" or "pro*" or "inv*" or "met*")

2-3 データベースの中国語抄録収録率

Orbit.com については、原語収録確認用コマンド(OAB=YES)を用いたが、PatBase、Discover については、N=20 による目視確認で、中国語抄録の有無を確認した。

2-4 出願人検索

中国出願人の例として、乐凯集团第二胶片厂(SECOND FILM FACTORY OF LUCKY GROUP)を取り上げ、中国語名、英語名による検索を実施し、中国公開特許のヒット件数、内容を比較した。

3. 検証結果

3-1 中国公開特許 英語抄録収録

(1) 1990～2010 年発行分

3種類の DB 共、概ね英語抄録は収録されている。なお、今回英語抄録収録は、人手翻訳データ又は機械翻訳データの少なくとも一方を含んでいるものを対象とした。(図.3)

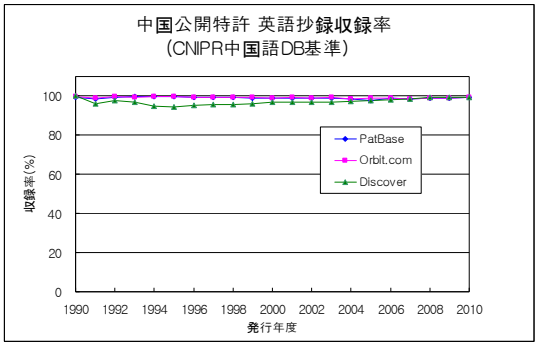

図 3. 英語抄録収録率

(2) 2010～2011 年発行分

収録タイムラグ確認のため、2010 年 6 月から 1 ヶ月毎の英語抄録率も確認した。(2011.8.25 最終確認) (図.4)

PatBase、Orbit.com の最新分は機械翻訳ながら、検索可能でタイムラグは少ない。 これらの DB の機械翻訳データは、表示のみでなく検索も可能である。

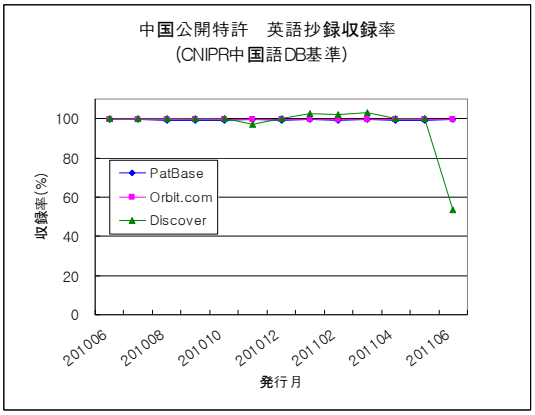

図.4　英語抄録収録タイムラグ確認

3-2 中国公開特許　中国語抄録収録

中国語抄録の有無についても、英語抄録と同様に確認した。
PatBase,Discover については目視確認のため検証精度が低いが、3 種類の DB とも中国語の収録は実務上問題ないレベルである。(図.5)

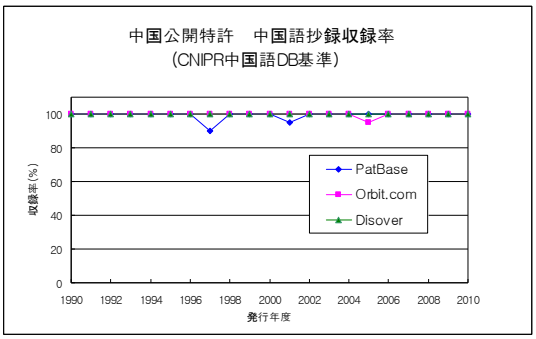

図.5　中国語抄録収録

ただし、発行後 1 年以内のデータについては、Discover で中国語抄録がないものが散見される。(図.6)

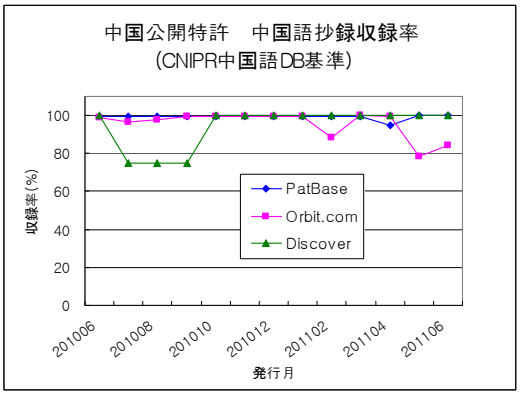

図.6　中国語抄録タイムラグ確認

3-3　出願人名の英語、中国語検索でのハイブリッド検索の効果確認

出願人:乐凯集团第二胶片厂
(特許英語表記例: SECOND FILM FACTORY OF LUCKY)について、英語単独及び、中国語併用での、網羅性について確認した。(図.7)
なお、英語出願人名検索については、(SECOND or 2)　and (FILM) and (LUCKY)　で確認した。
この結果、3DB 共、英語出願人名検索に比較して、ハイブリッド検索では網羅性が向上し、特に PatBase で網羅性の向上が大きく見られた。

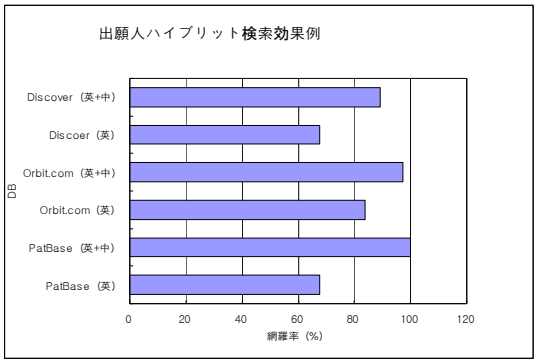

図.7　ハイブリッド検索の効果の確認

網羅性の向上が最も大きかった PatBase について、英語出願人名解析を行ってみると、英語出願人名では、表.1 のように表記割れを起こし、かつ

LUCKY にあたる部分の中国語の読み(ピンイン)からくる LEIKAI という表記もあることが判明した。

表.1　英語出願人名解析(PatBase)

乐凯集团第二胶片厂　　表記割れ例	率（%）
SECOND FILM FACTORY OF LUCKY G	60
SECOND FILM FACTORY OF LUCKY GROUP	12
NO 2 FILM FACTORY LEKAI GROUP	10
NO 2 FILM FACTORY LUCKY GROUP	7
LEKAI GROUP NO 2 FILM FACTORY	5
NO 2 FILM FACTORY OF LEKAI GRO	2
NO 2 FILM FACTORY LUCKY GORUP	1

また表記割れ以外にも、英語出願人情報の収録タイムラグ(4ヶ月以上)もあり、英語出願人名検索のみでは、中国特許検索の網羅性を上げることは難しい。

更に Orbit.com,Discover のハイブリッド検索でも、検索もれ原因を確認したが、特許の収録タイムラグに起因するもれが見受けられた。

4. 考察

3 つの英語+中国語キーワードのハイブリッド検索可能な DB について、収録を調べた結果、中国語の抄録が英語の抄録と同等以上の収録があることが確認できた。ただし、100%収録までにはタイムラグ及び、データの抜けもあり、データ収録の強化をお願いしたい。

また、 出願人名検索でハイブリッド検索の検証を実施したが、英語のみの検索より網羅性があり、CNIPR 中国語 DB 同等レベルに達するが、DB によっては収録タイムラグがある場合もあり、注意が必要である。

なお、中国語特許の検索においては、収録されているクレーム以下のデータの入手方法が非常に重要で、収録率、タイムラグ、データ精度に大きく影響する。確認したところでは、以下の①～③のような、大きく 3 つのデータ入手ルートがあり、それぞれ一長一短がある。検索漏れを防止するために、使用している DB の特色を把握した上で、調査する必要があると考える。

①IPPH(中国知識産権出版社)からデジタルテキストデータ購入
　(Orbit.com,Discover など)
②独自 OCR 処理 DB
　(PatBase,Dialog File325　など)
③独自翻訳　(WPI など)

なお、ハイブリッド検索では、中国語で検索した場合でも、英語(機械翻訳を含む)で表示され、1 つの DB で完結することから、効率的に調査でき、機械翻訳などによる誤訳の回避、及び英語情報収録のタイムラグ緩和策として有効で、今後の DB の一つのトレンドとなると考える。

本報告は、「アジア特許情報研究会」の2011年度の研究内容の一部を報告するものである。

5. 参考文献

[1]田畑文也　他:新 CNIPR の機能を検証する
(第 7 回情報プロフェショナルシンポジウム, 2010/11)
[2]伊藤徹男　他:中国・台湾および韓国特許庁データベースの全文検索機能とその活用
(第 6 回情報プロフェショナルシンポジウム, 2009/11)