アジア・新興国特許調査における原語抽出法

○伊藤徹男[1)]
アジア特許情報研究会[1)]
〒300-1260　茨城県つくば市西大井1733－15
Tel: 090-8700-7256
E-mail: patentsearch2006@yahoo.co.jp

Source language extraction method in Asian and emerging country patent search.

ITO Tetsuo[1)]
Society of Asia Patent Information[1)]
1733-15 nishiooi Tsukuba-city Ibaraki Japan
Phone: +81-90-8700-7256
E-mail: patentsearch2006@yahoo.co.jp

【発表概要】

商用データベースへのアジアや新興国の特許情報収録が不充分な中でWIPOのPATENTSCOPEやASEAN PATENTSCOPEには、各国特許庁が有する特許情報が収録され、利用できるようになった。これまで、商用データベースにおいては収録国も限定的で、また収録言語も主に英語情報であった。PATENTSCOPEには、中国や韓国など東アジアやASEANその他新興国の特許情報が原語ではあるが収録されるに至った。その内容の詳細は、別途「アジア・新興国特許調査における無料データベースの実力検証」という形で報告予定であるが、本発表ではこれら原語データベースを検索するに当たり必要となる「原語」をどのように抽出するかを紹介する。

商用英語データベースにおいては、中国特許情報なども書誌・要約だけなく、請求の範囲や全文まで機械翻訳ないしは人間翻訳による英語情報が収録されるようになったが、英語情報には誤字・脱字だけでなく誤訳も存在し、調査担当者も英語情報を補完する目的で各国特許庁データベースにアクセスして原語検索や査読をするようになりつつある。

もちろん、若干の検索漏れなどが許される出願前調査や先行技術情報の把握などでは、日本語や英語で検索できる（サーチャーにはフレンドリーな）システムを使うことで充分な場合もあるが、当該国で事業展開を図る場合の権利侵害調査や無効化資料調査においては、網羅的な調査が求められるので、機械翻訳などによる日本語や英語での調査では充分とは言えない場合もある。

現状では、多くの調査担当者（サーチャー）は英語以外の各国原語の読み書きができないと思われるので、そのような各国原語を理解できない状況の中で、どのように原語を抽出し、検索式を立てればよいかの指針となれば幸いである。

【キーワード】
アジア, 新興国, 特許調査, 原語検索, 原語抽出, 誤訳, PATENTSCOPE

1. はじめに

中国や韓国など東アジアの国に加え、ASEAN など新興国の特許情報がデータベースとして整備されてきた。

東アジアの特許情報が商用英語データベースに搭載され、登録情報も含め完備されたのは 2010 年以降である。それ以前は、中国、韓国の英語データベースには登録情報がなかった。特に、現在でも韓国では、公開公報より前に登録公報が発行されると公開公報が発行されないので、公開公報だけの調査では漏れが生じる状態である。

それに加え、DOCDB(EPO が管理する worldwide の書誌データベース)を元とする商用英語データベース、中でも中国特許の英語情報には誤字・脱字などのスペルミスの他、誤訳などの問題から英語情報だけには頼れない（検索漏れが生じる）という状況が続いている。

また、ASEAN など新興国の情報を収録する商用英語データベースも限られており、英語情報では充分な調査ができないのが現状である。

そのような中、PATENTSCOPE には 2017 年以降、Google Patents には 2018 年以降、東アジア、新興国を含めたワールドワイドな情報が搭載されるようになった。PATENTSCOPE へ収録された ASEAN 特許情報の概要は既に紹介されているが [1), 2)]、さらに詳細を本シンポジウムの別テーマでも紹介予定である。

PATENTSCOPE に収録された各国情報の元データは各国特許庁データベースに収録されている原語情報であるので ASEAN ではインドネシア、タイ、ベトナムについてはそれぞれの各国原語での検索、査読が必要となる。

したがって、東アジア、ASEAN の特許調査では英語データベース情報を補完する目的で原語調査することが必要になるが、多くの日本人サーチャーは各国原語での調査には不慣れな場合が多い。

そこで、各国原語に精通していないサーチャーのために英語以外の各国原語の抽出法および各国原語情報の査読についての一方法を紹介する。

2. 英語データベースのスペルミス、誤訳の現状

商用英語データベースの多くは各国特許庁からの情報を元にした DOCDB をベースとする情報であるが、特に中国特許英語情報のスペルミス、誤訳が問題となっている。

例えば、ポリエチレン(polyethylene)については、「polethylene」（脱字）や「ployethylene」（誤字）が存在するが、これらは「polyethylene」では検索できないが、査読では polyethylene と読める。しかし、ポリエチレンを表す「聚苯乙烯」または「聚苯烯」などの表記がないにもかかわらず、聚乙烯(polyethylene)が「polystyrene」などと誤訳されている場合には対応できない。

このような英語のスペルミスや誤訳も「聚乙烯」と中国語で検索すれば「ポリエチレン」を問題なく抽出できる。

出願人検索においても「Fuji Fiml」「Mutata Manufacturing」などのスペルミスや「马渊马达」(Mabuchi Motor)が「Mazda Motor」と誤訳されている例など挙げればきりがないが、これらも中国語検索によって目的のものを抽出できる。

また、中国特許英語データベース中にはドイツ語、フランス語表記も存在し、これらも多くの場合、英語では検索できないので注意が必要である。

3. 原語検索の必要性と原語の抽出

このように英語データベースではスペルミスや誤訳などによる検索漏れが発生するために東アジアや ASEAN など新興国の特許調査では原語検索での補完が必要となる。

もちろん、原語データベースから原語で検索し、原語で査読することも可能であるが、中国特許や韓国特許においては日本特許庁が提供する「中韓文献翻訳・検索システム」（以下、JPO 中韓システムという）によって日本語で査読可能であり、査読の効率も英語情報より良好なことは周知のことである。

しかし、原語検索の必要性があるとしても中国、台湾、韓国など東アジアだけでなく、インドネシア、タイ、ベトナムなどの日本人にとってなじみのない原語をどのように抽出すればよいかが問題となる。

そこで、原語抽出用ツールとして、まず挙げられるのは Google 翻訳ツールではないかと思われる。しかし、他の翻訳ツールと同様にこのような翻訳ツールでは1つの用語に対し、1 つの訳語しか表示してくれず異表記まで網羅的に表示できるツールはほとんどない。

Google 翻訳では、
Three dimensional printer
⇒ 三维打印机
3D printer ⇒ 3D 打印机

PATENTSCOPE には・CLIR（多言語検索）機能があり、中国語や韓国語に対応し、用語の訳語と類義語を抽出して検索できる、とされているが、やはり異表記を網羅的に抽出するものではないし、不正確な用語も抽出されるので利用できない。

今回紹介する各原語抽出用のツールはいずれも無料のデータベースである。中国語（簡体字）では Innojoy(http://www.innojoy. com/search/)、中国語（繁体字）では台湾特許庁の TWPAT (https://twpat6. tipo.gov.tw/)、韓国語では韓国特許情報院の KIPRIS(http://engpat.kipris.or.kr/engpat/searchLogina.do?next=MainSearch)であり、インドネシア語、タイ語、ベトナム語については、Google 翻訳ツールと各国特許庁データベースを組合せて抽出する。

予稿においては紙数の関係から中国語（簡体字）の例のみを挙げる。

いずれも英語（異表記）から原語を抽出する方法である。データベースからの抽出で確認が容易なように、まずは「発明の名称」から検索して抽出する。

【中国語（簡体字）異表記の抽出】

Innojoy コマンド検索で TI=(three dimension% print% or 3D print%)と検索すると 10 件ほどの発明の名称に「3D 打印」という中国語を確認できる。

次いで、TI=((three dimension% print% or 3D print%) not (3D 打印))と検索すると「3D 印刷、三维打印、三维印刷、三维列印」などその他の異表記を確認できる。これらの異表記を not 用語に加えて再検索する、というように繰り返し検索して検索結果がゼロになるまで検索することで異表記を網羅することができる。

このようにして発明の名称から抽出した異表記は、「三维打印,三维列印,三维印刷,三维印花,三维印相,三维印录,三维造型,立体打印,立体列印,立体印刷,立体印花,积层打印,3D 打印,3D打印」などとなるが、「三维（3D）打印」「3D 电脑

打印」「3D 可打印」などや、同様に要約、請求項中の異表記なども考慮して、「Three dimension or 3D」として「三维,立体,积层,3D,3D」、「print」として「打印,列印,印花,印表,印刷,印相,晒印,造型」を近接演算することで、かなり網羅的に抽出することができる。

また、3D print%では、3D,3Dだけでも半角、全角を組合せた「3D 打印,3D 打印,3D 打印,3D打印,三 D 打印,3D 列印,3D 印刷」の表記もあり、近接演算の 3D として「3D,3D,三 D」も加える必要がある。

以上から、以下のような式ができる。

TI,ABST,CLM+=((三维 or 立体 or 积层 or 3D or 3D or 3D or 3D or 三 D) pre/3 (打印 or 列印 or 印花 or 印表 or 印刷 or 印相 or 晒印 or 造型))

中国語の出願人異表記の抽出も上記同様、英語出願人名から抽出する。

PA=(FUJI FILM or FUJIFILM or FUJI PHOTO FILM)からは以下が抽出されるが、グローバルな会社では「富士胶片株式会社」のように法人格まで付けて検索しないと関連会社まで抽出することになる。

本体と思われるもの
- 富士胶片株式会社
- 富士写真胶片
- 富士写真软片
- 富士摄影胶片
- 富士胶片摄影

関連会社
- 富士胶片和光纯药
- 富士胶片富山化学
- 富士胶片电子材料美国
- 富士胶片制造欧洲
- ・・・

用語も出願人名も中国特許情報では異表記も多いので工数は多くなるが、網羅的な検索を行いたい場合には、多少の工数の増加はやむを得ないと思われる。

4. おわりに

英語表記のスペルミスや誤訳などが多い、また異表記も多い中国特許情報についての原語抽出法を紹介したが、ASEAN 各国の原語抽出法はもう少し簡単である。

英語（異表記）を元に not 演算して求める、という点は共通であり、それらについてはシンポジウムの場でそれぞれ紹介したい。

本手法については、これまで各種のセミナーにおいて紹介してきた内容を一部含むが、セミナー情報（資料）は誰でも入手できる情報でもなく、引用されることもほとんどないので、ここに改めて紹介することとした。

5. 参考文献

[1] 伊藤：PATENTSCOPE による ASEAN 特許調査概要（Japio Year book2018）
http://www.japio.or.jp/00yearbook/files/2018book/18_2_04.pdf

[2] 「ASEAN における各国横断検索が可能な産業財産権データベースの調査報告」(JETRO 2018)
https://www.jetro.go.jp/ext_images/world/asia/asean/ip/pdf/search_ip_communique_asean2017.pdf

[3] 特許文献機械翻訳の品質評価手法に関する調査(Japio 2014)
https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/document/kikai_honyaku/h25_01.pdf