# 技術動向の把握と特許文書処理技術
## ―新たな特許マップの構築に向けて―

Patent Processing for Analysis of Technology Trend

東京大学名誉教授／マイクロソフトリサーチアジア研究所首席研究員 **辻井　潤一**

**PROFILE:** 国際機械翻訳協会（IAMT）およびアジア太平洋機械翻訳協会（AAMT）元会長、AAMT／Japio 特許翻訳研究会委員長、国際計算言語学会（ACL）元会長、国際計算言語学委員会（ICCL）永久メンバー

jtsujii@microsoft.com　　TEL +86-13520978845

## 1 はじめに

特許の情報から、特定の技術分野の動向、技術分野からみた特定企業の特徴を把握する技術は、テキストマイニングの典型的な応用として研究されてきた。特許申請の時系列的な変化、技術間の相関関係を把握し、視覚化技術で見やすくグラフ表示した特許マップの作成は、平成９年から平成 12 年にかけて特許庁の事業として実施され、その成果が刊行物とウェッブサイトで提供されている [1]。

この事業が実施された時点と比べると、計算機を使った文書処理、ビッグデータ処理のためのインフラ技術、使用できる文書集合など、技術動向の分析を支援する特許マップの技術環境は一変している。

計算機で処理可能な文書集合は、日本の特許庁が持つ日本語の特許情報だけではない。英語、中国語、韓国語、ドイツ語など、多言語の特許文書が比較的簡単に手に入る。しかも、ヨーロッパの特許庁が Google の機械翻訳システムの運用を始めているように、情報技術を積極的に活用し言語の壁を解消しようという動きが活発化している。日本の特許だけを分析対象にした特許マップではなく、世界の特許を対象とした特許マップ、国別や地域別のきめ細かな技術動向を分析するためのツールの開発が可能になりつつある。

対象文書の多様化は、特許文書の多言語化だけではない。技術に関するアカデミックな論文や報告書、新聞や雑誌記事など、特許文書だけでは得られない情報を含んだテキストも、電子的な流通が普遍化し、簡単に手に入る。豊富な技術情報も Wikipedia などのウェッブ文書で公開されている。特定企業の新しい製品と基盤技術の関係や、企業間の提携や競合関係、技術間の相互関係など、特許文書だけでは得られない情報が、これらの文書から取得できる。

前回の特許マップ事業の時点では、特許文書の内容を計算機で処理するための文書解析技術、あるいは、大量の情報項目間の相互関係を分析するビッグデータのマイニング技術、ユーザの視点からのデータ分析の結果を即座に視覚化する技術などは、まだ実用段階には達していなかった。また、構造化データベース（たとえば、企業の年度ごとの収益、製品の出荷数などを示す数値情報を蓄積したデータベース）からマイニングの結果を特許情報という文書情報と結び付ける技術、いわば、定量的なマイニングと定性的なマイニングを結び付ける技術も、未熟であった。

本稿では、前回の特許マップ構築当時から大きく変化した技術環境を前提として、次世代型の特許マップ構築に向けての議論を行う。特に、筆者が専門とする文書処理、知識 Linking の技術が次世代の特許マップにどのような寄与をするかについて考える。

## 2 情報抽出と知識 Linking

過去 10 年間に大きく進展を遂げた文書処理技術に情報抽出（Information Extraction）と知識 Linking の技術がある。次の２つの文を考えてみよう。

(1) [Skype] is a proprietary [voice-over-IP] service acquired by [MS].
(2) [MSFT] stock soars after acquiring [VoIP] company [SKPY].

この2つの文には、「マイクロソフトがスカイプを買収した」、「スカイプは VoIP の技術を持った会社である」という技術動向を分析するのに有効な2つの情報が含まれている。

表面上の表現の差を無視して、(1)(2)の2つの文からこの共通の情報を取り出すタスクを情報抽出という。人間にとっては簡単なこの作業は、計算機にとっては存外難しい。たとえば、(1)の MS は、この文中ではマイクロソフトを指すが、

(3) In MS, the insulating covers of nerve cells in the brain are damaged.

の MS は、多発性硬化症(multiple sclerosis)を指している(名前表現の**曖昧性**)。また、Voice-over-IP と VoIP, Skype と SKYPY, MS と MSFT のように、違った表現が同じ会社や技術を指すことも多い(名前表現の**多様性**)。

このように、名前表現の曖昧性、多様性の問題を解決して、テキスト中の表現(単語や句)を現実世界での特定の会社や技術と適切に結び付ける技術は、Entity-linking と呼ばれ、情報抽出や知識 linking の中核的な技術となっている。

個々の会社や特定の技術のように、一意に認識できる要素にユニークな識別番号が割り振ふり、その識別番号を使って、その要素に関する情報を組織化しようという試みは、Semantic Web の考えにもある。各種のデータベースや知識ベースが Entity に関する共通の識別番号を使って構築されていれば、テキスト中の名前表現に識別番号を割り振ることで、テキスト中の名前表現とデータベース中の Entity に関する「知識」とが結び付く。Entity-linking は、知識 -linking の第一歩である。

Entity に関する知識としては、特定企業の年度ごとの収益や特定製品の出荷数など、構造化されたデータベース中のデータ項目がある。さらに知識ベースが蓄積している「マイクロソフトの CEO が S. Ballmer」、「マイクロソフトの製品には、Surface、Office、Windows,。。。」といった事実(Fact)、「VoIP 技術は、VoPN(Voice-over-Packet Network)技術の一種」といった一般的な知識など、数値化できない情報もある。

事実や一般的な知識を計算機内に蓄積した知識ベースも、過去 10 年間の間に飛躍的に拡大している。実際、現在日常的に使われている Wikipedia の構築も 2001 年ごろから始まったものであり、前回の特許マップの事業時にはまだ姿すら現していなかった。さらに、現在では、テキスト情報を主体とした Wikipedia から一歩進んで、計算機処理用に構造化された知識・事実ベース(Freebase, Yago など)の構築とその利用も急速に進展している。これらの知識ベースでは、すでに個々の Entity には識別番号が割り振られている。Entity-linking は、テキスト中の表現とこれらの構造化された知識・事実ベースとを結び付ける基本的な技術である。

## 3 関係情報の抽出

「マイクロソフトがスカイプを買収した」、「スカイプは VoIP 技術を持つ」といった情報は、会社(マイクロソフト)と会社(スカイプ)、あるいは、会社(スカイプ)と技術(VoIP)という、2つの Entity 間の関係を示している。こういった関係の認識にも、Entity-linking の場合と同様な多様性と曖昧性の問題がある。たとえば、

(4) The HTC **Amaze 4G** runs the **Android operating system**.
(5) The **Android operating system** which is used in Smartphone devices such as the **HTC Amaze 4G** has become …

では、ある製品(Amaze 4G)がある技術(Android operating system)を使っているという関係が、表面上異なった表現で捉えられている(**多様性**)。 また、(4)では製品とその製品に使われている技術との関係を表している動詞 run は、

(6) **Ballmer** runs **MSFT** as CEO

という文では、人物とその人物が属する組織との関係を表現する（**曖昧性**）。

関係表現の多様性と曖昧性を解消し、標準的な関係の集合に写像するのが、Relation-linking 技術である。この技術は、テキストから新たな関係を認識したり、関係が記述されているテキストの特定箇所とその関係に関する情報が蓄積された知識・データベースの項目とをリンクすることができる。

特定の企業が持つ技術と製品、企業間の提携と競合、応用技術と基礎技術、特定企業とその企業が得意とする技術分野や苦手な技術分野、といった関係は、特許、新聞記事、Ｗｅｂドキュメントなど、多様な文書群から Relation-Linking の技術を使うことで取り出すことができる。

## 4 特許マップとテキスト処理、知識処理

10 年以上も前に実施された特許マップの事業では、特許情報にあらかじめ人手で付与されたメタデータを唯一の分析対象としていた。また、分析過程で使われた知識ベースも、IPC などの標準的な分類体系のみであった。分野間の相互関係、技術間の相互関係を IPC コードで捉え、特許件数の推移などを視覚化した。実際は、事業実施時点でのテキスト処理、知識処理の状況では、特許文書の内容自体を処理することは不可能であった。

また、視覚化のための道具立てもまだ十分整っていなかったことから、前回の特許マップでは静的なグラフによる静的な表示のみが与えられている。

解析結果を視覚化し、それに人間が関与して更なる分析と視覚化を行うという、人と計算機との共同作業をサポートする動的な視覚化技術は使われていない。現在、急速に進展しつつあるビッグデータの解析では、このようなインターラクティブなデータ解析と部視覚化技術は不可欠であり、格段の進歩を見せている。

さらに重要な状況の変化は、特許文書の集合体を分析して視覚化した特許マップではなく、特許文書、新聞記事、論文、技術報告書など多様な文書群や、文書情報以外のデータベースや知識ベースを有機的に結び付けることによる技術動向、製品動向、技術間の相互関係の分析が出来る技術的な基盤が整ってきたことである。

このようなさまざまな技術を一つのシステムとして纏め上げることで、技術動向分析、企業体や製品と技術との相互関係の分析を行い、知財戦略、ビジネス戦略、研究開発戦略などの立案するための、次世代型の特許マップシステムが構想できる。

## 5 多言語特許マップ

経済のグローバル化、技術開発の国際的な分散化にともない、知財管理や知財戦略における多言語情報の重要性が高まっている。

筆者は、アジア太平洋翻訳協会（AAMT）と JAPIO が共同して運営している AAMT/JAPIO 機械翻訳委員会の委員長をしている。この委員会が設立された 10 年前に比べると、韓国、中国、台湾など、アジア諸国の研究開発力が飛躍的に伸び、いまや特許情報の管理や分析におけるアジア諸言語による特許の取り扱いは不可欠になっている [2]。

膨大な数の特許文書、しかも、多言語化が急速に進展する状況に対処するためには、機械翻訳に代表される情報処理技術を使った効率化が不可欠である。上記の委員会でも、特許文書の機械翻訳、特に、中国語や韓国語の特許文書の翻訳を積極的に取り上げて、中国や韓国の対応する機関とも共同しながら、調査研究を行ってきた。

この委員会では、特に、特許翻訳における専門用語の取り扱いに焦点を当ててきた。特許のような科学技術に関連した翻訳では、科学技術の専門用語が誤り無く翻訳できることが非常に重要だと考えられるからである。

専門用語というのは、会社名や人名といった典型的な Entity をあらわす表現とはいえないが、前述の例に現れた VoIP のように、特定の技術や技術分野をあらわす表現である。このような専門用語は、同じ技術内や技術分野を英語、日本語、中国のいずれの言語で表現するかは、重要ではない。どの言語の専門用語を使っても、同じ技術や技術分野を指すことが専門家の間で合意できる。いわば、多言語による専門用語の差は、名前表現の

多様性の一種であると考えてよい。

言い換えると、人名、組織名が表現する Entity に識別子を与えて、同じ Entity を指し示す多様な言語表現をすべて同じ識別子にリンクするという考え方は、異なる言語における専門用語に対しても適用できる。この専門用語 の linking（すなわち、広義の Entity-linking）ができると、言語が異なっていても、同じ技術や技術間の同じ関係を記述している文書を同定できることになる。必ずしも、特許文書の翻訳は完全にできていなくても、国別の技術動向の比較、国の異なる企業間の技術能力の相互比較など、グローバル化の中の戦略立案に不可欠な情報分析をサポートするシステムが可能となる。

上記の委員会では、英語、中国語、日本語の特許文書から、専門用語の辞書を自動構築する技術の調査研究を行い、実際にシステムを構築している。このような研究の成果は、特許や技術文書の機械翻訳だけでなく、多言語文書からの特許マップの構築に活用することができると考えている。

## 6 おわりに

前回の特許マップ構築プロジェクトから 10 年以上の時間が経過した。その間、特許のような技術文書を取り扱う多言語処理技術、ビッグデータの解析とその視覚化の技術は、飛躍的に進歩した。また、新聞記事、科学技術論文、各種の技術報告書など、特許文書以外にも、技術動向の分析に有効な文書のほとんどが計算機を通して流通している。本稿では、この状況の激変が、前回の特許マップとはまったく質の異なるシステムの構想を可能にしていることを述べた。

知財の管理と戦略立案を正確、迅速、的確に行うための技術は、イノベーションが国家の衰亡を決定づける近未来の最重要技術である。日本の国として、この分野に積極的な投資が行われることを期待している。

**参考文献**

[1] 特許マップ事業のサイト http://www.jpo.go.jp/shiryou/s_sonota/tokumap.htm
[2] 日本特許情報機構：Japio Year Book 2012