特許と論文を対象にした技術動向分析

難波英嗣（広島市立大学大学院）
奥村学　（東京工業大学）
新森昭宏（インテック・ウェブ・アンド・ゲノム・インフォマティクス株式会社）
谷川英和（IRD 国際特許事務所）

はじめに

2004 年 7 月より，著者らは NEDO 産業技術研究助成事業の支援を受け，「特許，論文データベースを統合した検索環境および動向分析ツールの構築」に関する研究を行ってきた．本稿では，これまで著者らが開発してきたシステムの中で，特に特許と論文を対象にした技術動向分析システムの概要を説明する．

ある技術分野において，「どのような要素技術がいつ頃から使われているのか」という情報を網羅的に収集し整理することは，その分野の技術動向を概観するのに必要不可欠である．しかし，このような動向調査には多くの時間と労力を要する．そこで本研究では，特許と論文データベースから技術動向情報を自動的に抽出し，可視化するシステムの構築を行う．

特許と論文を対象に技術動向情報を行うためには，特定の分野の文献を収集し，そこから技術動向情報を抽出する，という 2 つの課題を解く必要がある．本研究では，前者についてはシソーラスの自動構築，後者については文書構造に基づいた情報抽出によって解決する．

本稿の構成は以下のとおりである．2 章では，まず著者らが開発した技術動向分析システムの動作例を示す．このシステムの要素技術として，3 章ではシソーラスの自動構築について，続く 4 章では文書構造に基づく技術動向情報の抽出について，それぞれ述べる．最後に 5 章で本稿をまとめる．

1.　技術動向分析システムの動作例

著者らが開発する技術動向分析システムは，以下の 2 点について調査することが可能である．

(1) ある分野で，どのような要素技術がいつ頃から使われてきたのか．
(2) ある要素技術がどのような分野で使われてきたのか．

図 1　技術動向分析システムのトップ画面

【図 1】は，システムのトップ画面で，(1)と(2)を調べるための検索フォームがそれぞれ用意されている．

図 2　「機械翻訳」分野の要素技術の一覧表示

【図 2】は，「機械翻訳」という用語をシステムに入力した時の解析結果を示している．【図 2】では，左端に「機械翻訳」の要素技術名が列挙され，各技術が使われた年が図の右側に示してある．例えば【図 2】の「構文解析」の場合，この技術を要素技術に用いた機械翻訳に関する論文が 1991 年に 1 件発表されており，これが図中で「●」として表示されている．ユーザが●上にカーソルを重ねると，その論文の書誌情報がポップアップ表示される．図では，ポップアップ表示として「ナレート ペッチャラニン / 田中清 / 中村康弘 / 松井甲子雄，タイ日機械翻訳のためのタイ語の構文解析，1991，全国大会，(情報処理学会)」が例示されている．

図 3　「HMM」を要素技術に用いている分野一覧

【図 3】は，このシステムにおいて，要素技術名「HMM」をフォーム入力した時の検索結果の画面である．これは，HMM が使われた分野一覧を示しており，画面の左側が HMM を用いた論文を，右側が特許を示している．この出力を見ると，HMM は学術界(論文中)では 1980 年代後半に音声認識分野で使われはじめ，その後，手書き文字認識や指紋照合等の画像認識分野，形態素解析や言語獲得等の自然言語処理分野でも使われるようになってきているのに対し，産業界(特許中)でも音声認識や画像認識では同様に HMM が要素技術として使われているものの，自然言語処理分野では使われた例が見当たらない，といった分析ができる．

また，本システムでは，シソーラスを用いて，「機械翻訳」の関連分野の技術動向も調べることができる．【図 2】において，画面上部の「シソーラス(機械翻訳)」というリンクをクリックすると，シソーラス(【図 4】)が表示される．【図 4】では，「機械翻訳」と関連のある様々な用語が表示されている．例えば，機械翻訳は「英語」「原文」等を入力とし，「日本語」「(第 2 言語)」等を出力とすることや，上位語として「自然言語処理」や「自然言語処理システム」があることが図に示され

ている.ユーザは,図中の用語をクリックすることで,今度はその用語を中心とした関連する様々な情報が閲覧可能である．また，画面上部の「動向分析」というリンクをクリックすれば,【図 2】や【図 3】のような技術動向分析の画面に戻ることができる.

図 4　「機械翻訳」の入出力 / 上位 / 下位語表示

2.　特許用語シソーラスの自動構築

シソーラスは，文献を検索したり，特許や論文等の専門文書を執筆したりする上で有用な情報源として活用されている．例えば，科学技術振興機構(JST)が提供する文献検索サービス JDReamII[1]では,ユーザが検索を行う際の支援機能のひとつとして，シソーラスがある．また，シソーラスは，情報検索や機械翻訳など計算機で言語処理を行う際の知識源としてもしばしば利用されている.しかし,シソーラスを人手で構築し,更新することは非常にコストがかかるため，テキストデータベースから，シソーラスを自動的に構築するという研究が近年活発に行われるようになってきている.

著者らは，特許データベースから，次に示す関係の用語を自動抽出した.

- 用語の上位，下位関係
- 同義語
- 入力 / 出力用語

以下,その抽出手法,および抽出結果について述べる.

3.1 用語の上位，下位関係の抽出

用語の上位，下位関係を抽出する代表的な手法は,「A や B などの C」や“A such as B, C”などの定型表現に着目したものである[Hearst 1992, 安藤 2003, 相澤 2006]．特に，特許の場合，明細書中でこのような定型表現が多用されるため，上位，下位関係の抽出に適していると考えられる．そこで，本研究でも，定型表現に着目した手法を用いる.

まず,「などの」「等の」「といった」「のような」の 4 種類の定型表現に着目し，公開公報(1993～2002 年)から，これらの表現を含む文を収集した．その結果,「などの」「等の」を含む文が 29,641,887 文,「といった」を含む文が 844,790 文,「のような」を含む文が 9,725,720 文収集された．実際に収集された文を見ると,「のような」と「といった」を含む文にはノイズが多く含まれていることがわかった．また，抽出された文数も「などの」と「等の」を含む文数と比べると件数は少なかったため,「などの」と「等の」の 2 つだけで十分な量の上位，下位概念が獲得できると判断した.【表 1】に，実際に獲得した上位，下位概念の概要を示す.

表 1　獲得された上位，下位概念

上位，下位関係(異なり数)	7,031,159
全用語数(異なり数)	1,825,518
1 語以上下位語を持つ用語数	833,215
1 語以上上位語を持つ用語数	1,236,663

抽出結果を調べたところ,若干の問題はあるものの,ある程度実用に耐えうる精度で抽出できている[難波

[1] http://pr.jst.go.jp/jdream2/

2007]. 例えば「ダイヤモンド」という用語に関して，次のものが抽出されている．

上位語：砥粒(988)，硬質材料(149)，宝石(128)
下位語：カーボン層(28)，カーボンダイヤモンド(27)

ここで，各用語に併記されている数字は，10 年分の公開公報中での出現頻度である．例えば，「ダイヤモンド」の場合，「ダイヤモンド等の砥粒」や「ダイヤモンドや○○などの砥粒」といった表現が 988 回出現していることを示す．

特許用語の上位，下位関係は，特許検索に有用であるばかりでなく，明細書を作成する際の支援ツールとしても利用できる．例えば，抽出された関係を用いて，「フロッピーディスク」の上位語を調べると「磁気記録装置」や「リムーバブル記録媒体」といった用語が得られる．請求項を記述する際，その請求項の権利範囲を広げるため，なるべく一般的な用語を使う必要があるが，「フロッピーディスク」の場合，「磁気で記録する」という仕組みに着目するのか，「持ち運べる」(リムーバブル)という機能に着目するのかによって，使う用語が変わってくる．

この他にも，例えば，「物を切断するのにどのような方法があるのか，可能な限り知りたい」といった用途にも使える．この場合，「切断手段」という用語の下位語を調べれば，「カッター」や「ハサミ」といった一般的なものから「ミシン目」「ウォータージェット」「レーザー」「加熱線条」「放電加工機」等，様々な切断方法が存在することがわかる．

3.2 同義語の抽出

「文書編集装置」には「文書作成装置」という同義語が存在する，という情報は，網羅性が重視される特許検索において必要不可欠なものである．しかしながら，幅広い分野にわたって同義語辞書を作成・保守しつづけることは，非常にコストがかかる．そこで，著者らは，2 種類の方法による同義語の自動抽出を試みている．

3.2.1 用語の上位，下位関係を用いた同義語の抽出

著者らが提案する手法は，用語の上位，下位関係を使って自動抽出するものである．【図 5】は，「文書編集装置」と「文書作成装置」という 2 つの用語を中心に，これらと上位，下位関係にある用語の一部を示している．図のように，「文書編集装置」と「文書作成装置」が似たような意味を持っているのであれば，数多くの上位語あるいは下位語を共通に持つと考えられる．

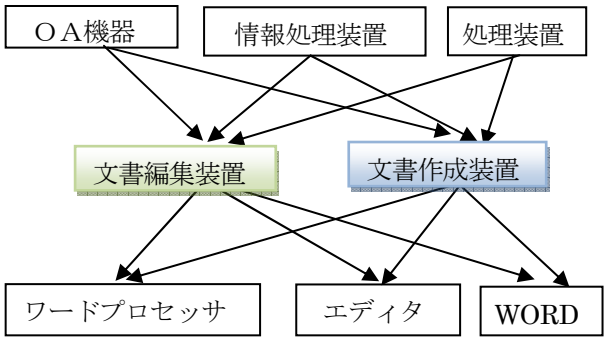

図 5　用語間の上位，下位関係を用いた同義語の検出

上で述べたアイディアは，引用分析研究における書誌結合[Kessler 1963]と共引用分析[Small 1973]に基づく．引用分析とは，論文間の引用・被引用関係を用いて，論文間の関係を分析する方法であるが，書誌結合は，論文間の関連度を測る時に，2 論文間でどれだけ同じ論文を引用しているか，他方，共引用分析は，2 論文がどれだけ他の論文で共に引用されているか，という基準を手がかりに分析を行う手法である．本提案手法は，用語間の上位，下位関係を論文間の引用関係と見なし，引用分析手法を用いて，同義語の抽出を行うものである．

実際に，書誌結合および共引用分析を用いて同義語

の抽出を試みたところ，どちらの手法でも同義語が抽出できていたものの，「アルミニウム」と「鉄」，「赤」と「緑」など，共通の上位概念または下位概念をもつ兄弟関係にある用語対が，同義語対と共に数多く抽出された．そこで，引用分析の結果から兄弟関係にある用語対の除去を試みた．

本研究では「A や B などの C」という定型表現に着目して上位，下位概念を抽出しているが，この表現において，A と B は兄弟関係にあると考えられる．そこで，定型表現から兄弟関係にある用語対を抽出しておき，引用分析の結果から，兄弟関係にある用語対を除去すれば，効率的に同義語の抽出ができると考えられる．

実際に，上記の定型表現から兄弟関係にある用語対を抽出したところ，5,046,426 個の用語対が得られた．この用語対を用い，引用分析結果の中から，兄弟関係にない用語のみを抽出した結果の一部(上位 15 件)が【表 2】に示されている．

表 2　同義語の抽出結果

順位	抽出された用語対	
1	ヒータ	ヒーター
2	医薬	医薬品
3	コントロールユニット	制御ユニット
4	伝導材料	伝導体
5	合穴	固定部
6	昇圧回路	高圧回路
7	不凍液	ブライン
8	デジタル表示	グラフ表示
9	電圧検出回路	波形整形回路
10	宝石類	宝飾品
11	樹脂液	感光液
12	偏光素子	複屈折板
13	ハロゲン化銀	銀塩
14	セラミックス	セラミック
15	活性溶媒中	溶媒中

3.2.2 特許間の引用関係を用いた同義語の抽出

「文書編集装置」と「文書作成装置」，あるいは「磁気記憶装置」と「磁気記録装置」といった同義語は，通常，同一特許中に出現することはほとんどない．しかしながら引用・被引用関係にある二つの特許 A と B において，特許 A に「文書編集装置」が，特許 B に「文書作成装置」が出現するというケースはしばしば見られる．そこで，特許間の引用関係に着目し，ある用語 X の同義語を以下の 3 つの手順で抽出する．

(手順 1) 用語 X と関連のある特許を収集する．

(手順 2) 手順 1 で収集された特許と引用関係にある特許を収集する．

(手順 3) 手順 2 で収集された各特許から，トピック語を抽出し，それらを頻度順に並べて出力する．

ここで，手順 3 において，各特許からどのようにトピック語を抽出するのかが問題になる．文書中の重要語を抽出する手法としては tf*idf 等を用いるのが一般的であるが，本研究では，特許中の請求項に着目してトピック語を抽出する．請求項中で，「において」の直前の名詞句や「を特徴とする」の後の名詞句は，その特許のトピック語を示すことが多い [新森 2004]．例えば，【図 6】に示す請求項の場合，「において」の直前の名詞句「シフトレバー装置」と「を特徴とする」の後の名詞句「シフトレバー」「シフトロック装置」がそれに相当する．本研究ではこれらの用語をトピック語として抽出する．

実際に，1993 年～2002 年の 10 年間の公開公報から，522,810 個のトピック語を抽出した．次に用語毎に手順 1～3 を適用し，最終的に 257,459 個の同義語対を抽出した．【表 3】に，一例を示す．

車体に固定する筐体内に前後揺動体を車体前後方向へ回動可能に軸支し、該前後揺動体に揺動基部を車体左右方向へ回動可能に軸支し、該揺動基部に植設したシフトレバーを車体前後及び左右方向へ揺動させることにより筐体上面に形成したゲート部を移動し所望のレンジを選択して自動変速機を切替操作する**シフトレバー装置**において、前記揺動基部に一対の上下に離間した突起部を設け、一方の突起部に当接可能なP係止部と、他方の突起部に当接可能なN係止部を有する回転ロック体を前記筐体に回動可能に軸支するとともに、シフトレバーがPレンジ又はNレンジに移動したとき該回転ロック体を回動させるアクチュエータを前記筐体に固定したことを特徴とする**シフトレバー**の**シフトロック装置**。

図 6　請求項からのトピック語の抽出例

表 3　同義語対の例

頻度	用語 1	用語 2
75	磁気記憶装置	磁気記録媒体
37		磁気ディスク装置
18		磁気ヘッド
13		磁気記録装置
11		金属薄膜型磁気記録

3.3 入力 / 出力用語の抽出

用語の中には，その用語の直後に「する.」を加えることで動詞になるものがある. 例えば,「形態素解析」や「機械翻訳」といった用語に「する.」を加えると「形態素解析する.」や「機械翻訳する.」という表現が得られる. こうした用語の多くは, 何らかの入力があり, それを処理して新たなものを出力する用語であると考えられる[2].

[2] 例えば「形態素」という専門用語に「する.」を付け加えた「形態素する.」という表現は存在しない.「形態素」は入力や出力のある用語ではない.

ここで，このような文は「A ヲ B ニ C する.」という文構造になっている場合が多い. この時, A につく「ヲ」と B につく「ニ」を抽出すれば，それが C の入力と出力になっていると考えられる. 例えば, C が「機械翻訳」の場合,「ヲ」から「日本語文」や「文書」や「文字列」などが,「ニ」から「英語」などが抽出できる．同様に,「形態素解析する.」の場合,「ヲ」から「日本語文書」などが抽出できる.

なお,「ヲ」以外にも「カラ」からも入力情報が,「ニ」以外にも「マデ」や「へ」から出力情報が得られる. さらに, 専門用語に「する.」を付け加える以外にも「をする.」を付け加えた場合(例えば,「機械翻訳をする」)にも，同様に入出力情報が得られる.

3.　特許と論文を対象にした技術動向分析

4.1 論文中の要素技術の抽出

論文中の要素技術の抽出は，論文表題の構造を解析することで可能となる．多くの論文表題には「A に基づいた」や「B を用いた」などの表現が含まれる．この A や B には, ある技術を実現するための要素技術を示す用語が一般的に含まれている．そこで，論文表題を解析し，A や B から専門用語を抽出し，その論文の著作年を X 軸に, 抽出された用語を Y 軸にとることで, ある分野の動向を示すグラフを作成することができる.

4.2 特許中の要素技術の抽出

新森らは，言語的な手掛かりに基づいて，請求項の構造を解析するツールを構築している[新森 2004]. この研究では，請求項を意味的にまとまりのある複数の節に分割し，節間の関係を解析し，修辞構造木と呼ばれる木構造をくみ上げることで，構造解析を行っている．節間の関係には Component や Precondition や Elaboration など 6 種類が設定されている. このうち,

Componentとは、「AとBとCを備えたX」という形式で記述されるもので、Xの構成要素がAとBとCであることを示す．本研究では，新森らのツールを用い，要素技術を抽出している．例えば，【図7】の請求項が入力されると，下線部が要素技術として抽出される．

操作手段によりアクチュエータを駆動して所望の作業を行う作業機において、前記作業の作業機構に作成する負荷を検出する負荷検出手段と、この負荷検出手段の検出値に応じた周波数の信号を出力する第１の周波数変換器と、当該負荷検出手段の検出値に応じた周波数のパルスを出力する第２の周波数変換器と、前記第１の周波数変換器から出力される信号を前記第２の周波数変換器からのパルスの出力期間だけ間欠的に出力する変調手段と、この変調手段の出力に応じて振動を発生する振動発生手段とを設けたことを特徴とする作業機の操作用仮想振動生成装置

図7　特許請求項の例

ここで，請求項のComponentは，【図7】のように長い名詞句として記述されることが多い．そこで，与えられた要素技術名が請求項のComponentに含まれている場合，特許中でその要素技術が用いられていると見なし，要素技術に用いている分野一覧(【図 3】)に表示する．

4. おわりに

本稿では，特許と論文を対象にした技術動向分析システムについて述べた．特許中の要素技術の抽出について，4.2 節でも述べたとおり，要素技術が長い名詞句であるため，現状では論文を対象にした技術動向分析との統合が十分であるとは言えない．今後は，この点の改良に取り組む．

謝辞

本研究で用いた特許データ(1993 年～2002 年の公開公報)および論文データは，国立情報学研究所の許可を得て，NTCIR テストコレクションを利用させていただいた．本研究は，NEDO 産業技術研究助成事業の支援を受けて行われた．

参考文献

［相澤 2006］相澤彰子 “類語関係抽出タスクにおけるコーパス規模拡大の影響”情報処理学会研究報告 自然言語処理，NL-175，pp.91-98，2006.

［安藤 2003］安藤まや，関根聡，石崎俊 “定型表現を利用した新聞記事からの下位概念単語の自動抽出”情報処理学会研究報告 自然言語処理，NL-157，pp.77-82，2003.

[Hearst 1992] Hearst, M.A. “Automatic Acquisition of Hyponyms from Large Text Corpora” Proceedings of the 14th International Conference on Computational Linguistics, pp.539-545, 1992.

[Kessler 1963] Kessler, M.M. “Bibliographic Coupling between Scientific Papers” American Documentation, Vol.14, No.1, pp.10-25, 1963.

[難波 2007] 難波英嗣, 奥村学, 新森昭宏, 谷川英和, 鈴木泰山“特許データベースからのシソーラスの自動構築”言語処理学会 第 13 回年次大会，pp.1113-1116，2007.

[Small 1973] Small, H., “Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents” Journal of the American Society for Information Science, Vol.24, pp.265-269, 1973.

［新森 2004］新森昭宏, 奥村学, 丸川雄三, 岩山真 “手がかり句を用いた特許請求項の構造解析”情報処理学会論文誌，Vol.45，No.3，pp.891-905，2004.