特許の引用情報にみられる論文情報の定量的分析のためのシステム開発

小林義英† 落合圭† 橋本定幸†
塩尻栄美子† 山崎雅和† 栗原正昭†
浜中寿† 坂内悟† 國谷実† 治部眞里†

特許情報および論文情報を格納したデータベースを整備し，特許の引用情報に含まれる論文情報を定量的に分析することを目的としたシステムの開発を行った．開発においては，特許情報と論文情報の膨大なテキストデータを扱うことによる課題があった．データクリーニングや名寄せなど，開発に際し発生した課題とその対応を中心に開発の詳細について示す．

Development for quantitative analysis system of journal articles in patent citations

Yoshihide Kobayashi† Kei Ochiai† Sadayuki Hashimoto† Emiko Shiojiri† Masakazu Yamazaki† Masaaki Kurihara† Hisashi Hamanaka† Satoru Bannai† Minoru Kuniya† and Mari Jibu†

We maintained the database that stored patent information and thesis information, and developed the system to make qualitative analysis of thesis information included in patent information as quotation. For developing the system, we faced various problems due to treating huge text data of patent information and thesis information. We describe the details and countermeasure about the problems occurred in the development process such as data cleaning and name identification.

1. はじめに

イノベーション創造推進事業体として，独立行政法人科学技術振興機構（以下，JSTとする）は，その中期目標において，以下のような政策的要請を受けている．

- 機構は，研究内容，研究成果に係る論文発表，口頭発表，特許出願の状況および成果の社会・経済への波及効果等について把握し，わかりやすく社会に向けて情報発信する．
- 機構は，本事業における研究が国際的に高い水準にあることを目指す．その指標として，論文被引用回数，国際的な科学賞の受賞数，招待講演数等の定量的指標を活用する．

これを受け，過去に支援した研究課題の社会・経済への波及効果，あるいは定量的な指標を活用して，如何にJST事業がイノベーションの創出に寄与できたかを分析することは，独立行政法人としての責務といえよう．

引用情報を含む文献データベースを活用することにより，論文と特許の内容の可視化分析，論文・特許の引用・被引用関係の分析（サイエンスリンケージ分析）などを行うことが可能となり，従来困難であった網羅的な波及効果分析において必要となる指標が得られる可能性がある．特許の引用情報にみられる非特許文献（特許以外の論文や図書などの文献，主として科学技術文献）は，科学研究が産業技術に与えた影響を定量的に特徴づけるものと考えることが可能であり，様々な角度で分析を行う価値があると推測できる．しかしこの観点で実施された分析は少ない．有力特許に引用された論文の計量書誌学的分析において，大学，公的機関，企業などのセクタごとのサイエンスリンケージの分析はあるが，研究者別，機関別のサイエンスリンケージの研究はほとんど行われていない．これは大量の論文および特許について，研究者名や機関名の名寄せが困難であったことが原因の一つにあると思われる．

そこで本プロジェクトでは，論文情報および特許情報のデータベースについて名寄せ等データの整理を行い，特許の引用情報にみられる論文情報を定量的に分析することを目的としたシステムの開発を行った．

2. 開発手順

本プロジェクトに際し，下記手順にて作業を進めることにした．

- 論文情報および特許情報データベースの構築
- 機関および研究者の名寄せ
- 特許の引用情報にみられる非特許情報と論文情報のマッチング
- 論文情報および特許情報の抽出用検索システムの構築

† 独立行政法人科学技術振興機構
Japan Science and Technology Agency

以下，各手順の作業内容と発生した課題について述べる．

3. 論文情報および特許情報データベースの構築

3.1 データベースの選定

特許に引用されている論文情報を定量的に分析するためには，論文の書誌を含む引用情報を有する大量の特許情報と，国内外を問わず大量の論文情報を入手し，特許の引用情報に記載された文字列と論文情報を参照できる形まで整備する必要がある．

本プロジェクトにおいては，特許情報として欧州特許庁が提供している PATSTAT（Worldwide Patent Statistical Database）を，論文情報としてエルゼビア社が提供している SCOPUS カスタムデータをそれぞれ用いることにした．

PATSTAT（以下，特許データベースとする）は世界 80 ヶ国から収集された 5,000 万以上もの出願情報を有している．リレーショナルデータベースに導入しやすい形で提供され，欧州特許庁の書誌データベース（DocDB）の一部として，出願番号，公開番号，優先日や出願日，公開日，出願人や発明者の名前と住所，国コード，発明の名称，国際特許分類，ファミリー，要約，非特許情報を含む引用文献情報などが記録されている[1]．特許研究者がデータベースとして使い，学会もあることからノウハウの蓄積も多いため，分析対象として選択した．

SCOPUS カスタムデータ（以下，文献データベースとする）は世界 5,000 以上の出版社の 18,000 誌以上のジャーナルを収録しており，3,800 万件以上の書誌・抄録レコードが存在している．また，1996 年以降は出版された論文には引用文献情報も存在している．データ量が多いことと，研究者と所属機関の対応関係が明確になっていることから，分析対象として選択した．また，特許データベースは英語であり今回特許情報と論文情報のマッチングを行う必要があるため，英語を使用していることも選択の大きな要因となった．

特許データベースも文献データベースもデータ規模が膨大となるためデータの絞込を行った．まずそれぞれのデータベースについて年数（論文は発行年，特許は出願年）によるデータの絞り込みを行った．また，特許データベースについては，日米欧の三極特許庁と WIPO（World Intellectual Property Organization）に特許出願が集中していることから，出願国による絞り込みを行った．さらに，文献データベースについては今回論文情報のみを対象とするため，論文以外の文献情報（単行本など）については対象外とした．

3.2 データベース環境の構築

処理を行う環境については，大量のデータ保持と高速検索が実現できるよう，下記の構成とした．また，文献データベースと特許データベースについてそれぞれ処理が可能なように，下記構成マシンを 2 台用意した．

表 1 データベース環境

区分	名称，スペック	備考
データベース	MySQL5	
OS	Red Hat Enterprise Linux	
ハードディスク	3TB	RAID5,SAS,15krpm
メモリ	48GB	
CPU	2.83GHz	クアッドコア

開発にあたって，データ規模を考慮し，下記点を念頭において作業を行った．

- リソース管理
  処理の先頭でメモリ使用量を確認し，メモリ不足前に処理を切断するよう配慮した．

- データベース管理
  データベースとして使用している MySQL は，高速である分動作が不安定になることが多かったため，処理前に使用するテーブルのチェックを実施してから処理を開始するようにした．また，処理によっては MySQL のパラメータを調整し，より高速かつ安定的な検索が実施できるようにした．

4. 機関および研究者の名寄せ

特許に引用されている論文情報を定量的に分析するためには機関名および研究者名に存在する様々な表記揺れを解消する必要があるが，文献データベースおよび特許データベースそれぞれのデータがもつ特徴などにより，名寄せ処理には各種の問題点が発生した．以下にそれらの問題点と本プロジェクトにあたって行った対応について述べる．

4.1 機関名寄せ

(1) 機関名表記

機関名には表記ゆれが存在し，以下の種類があった．実際にはそれぞれの表記揺れ種類に加えて，表記揺れ種類が合わさったものなど様々な表記揺れが存在した．

表 2　機関名の表記ゆれ種別

種別	説明	例
正式名称	機関の正式英名	The University of Tokyo
略称	機関の正式英名の略称	JST
別表記	正式英名と異なる語順の表記	Tokyo University
一部略記	一部に略記が使われている機関名	Tokyo Univ.
機関内の組織名	機関内の部門名やプロジェクト名	CREST
誤記	機関名のスペルミス	The Univercity of Tokyo
旧機関名	機関の統合・分割などによる変更前の名称	Research Development Corpration of Japan

これらの表記ゆれを解消するため，同じ機関名を示す表記パターンについて同一の機関 ID をつけ，機関名辞書（機関名の各種表記パターンと機関 ID が組になったデータ）を作成し，それを元に名寄せを行うことにした．機関名辞書作成にあたっては JST の運営する科学技術文献検索サービス「JDream Ⅱ」の機関名辞書をベースにしたが，海外機関は整備が進んでいないためデータ数が少なく，国コードの情報も持っていなかった．また国内機関であっても，「JDreamII」の機関名辞書が JST で運用しているデータベースに合わせて作成されていること等から，特許データベースと文献データベースに現れる国内機関に関わる表記揺れの多くが「JDreamII」の機関名辞書から漏れていることが分かった．そこで本プロジェクトでは「JDream Ⅱ」の機関名辞書の中から特に今回注目した日本の主要大学および研究機関 152 機関（以下，名寄せ対象機関とする）について名寄せ処理の精度を向上させるため追加のデータ整備を行った．

また，特許データベースの出願人情報と発明者情報には機関名と研究者名を分ける区分はない．この点について検討した結果，本プロジェクトにおいては，出願人情報を機関名情報として，機関名名寄せの対象とした．つまり，ある特許の出願人の中にその機関が存在しなければ，発明人の中にその機関のデータが存在したとしても，その機関の特許と扱わないこととした．これは，特許の集計を行う際に，ある機関の特許として集計すべきデータは，その機関が出願人となっているケースが多いと推測したからである．

(2) 辞書管理

上述の通り 1 つの機関を表す表記パターンの数は非常に多いので，それを全て辞書に登録すること自体非常に労力がかかり，また後のマッチング処理において多大な負荷がかかる．そのため，機関名辞書に細かい表記パターンを増やすことはせず，マッチングの前処理として JST で所有する略記辞書による変換を実施することにした（略記の具体例については表 3 を参照）．機関名辞書に登録されている表記パターンと，名寄せ対象の機関名データの両方を，同じ略記辞書を使って変換し，その後両者をマッチングさせるようにした．

表 3　略記辞書の具体例

略記	略記前の元の表記
technol	Technolog technologia technologica technological technologice technologie technologies technologiques technology Thechnology
adv	advanc advance advanced advancement advancements advances Advancing

しかし，名寄せ対象機関のみ辞書の登録パターンを拡充させるだけでは名寄せの精度は向上しない．名寄せ対象機関以外の機関名データが名寄せ処理に障害を与えるためである．例えば東京大学と東京理科大学の場合，「Tokyo University of Science」は東京理科大学を表すが，もしこのパターンが辞書に登録されていなければ，後述するマッチング処理方式によって東京大学を表す「Tokyo University」に名寄せされてしまう．そのため，名寄せ対象機関についての名寄せの精度を下げないよう，名寄せ対象機関以外の機関名の表記パターンを人手で洗い出し，機関名辞書に登録した．

本プロジェクトでは辞書整備について，表記パターンの洗い出しと辞書登録の判断について多くを人手に頼ることになった．この部分については自動化できるよう検討を続けていく必要がある．

(3) マッチング処理

マッチング処理を行う際，辞書にある単語の使用順序は重要となる．例えば，東京理科大学を示す「Tokyo University of Science」という辞書パターンよりも先に，東京大学を示す「Tokyo University」という辞書パターンを処理対象としてしまうと，「Tokyo University of Science」という機関名も「Tokyo University」という辞書パターンに一致してしまうことになる．このようなことを防ぐために，辞書パターンの中で文字列の

長いものから順番にマッチング処理を実施した.

また，処理後のデータを検討したところ，1 つの機関名データが 1 つの機関を表している場合以外に，「Tokyo University and JST」のように複数の機関を含んでいるケースが存在した. これについては，両方の機関にマッチするように処理を行った.「Tokyo University and JST」の例では，まず「Tokyo University」と一致するので，一致した部分を除いて「and JST」という文字列に変換し，「Tokyo University」以外の辞書パターンから処理を続けることで，「JST」にも一致するようにした.

4.2 研究者名寄せ

(1) 研究者名表記

研究者名の表記においても先に機関名表記に述べたような問題があり，研究者全体についての名寄せを行うことは難しいため，名寄せ対象者を CREST や ERATO など JST がファンディングを行ったプロジェクトのリーダー格の人物に絞った（以下，名寄せ対象者とする）.

(2) 辞書管理

名寄せ対象者に対して，研究者の ID と所属機関の ID を組にした研究者名辞書を使用した. 後述する所属機関の判定処理は，研究者名辞書に登録されている所属機関 ID と，機関名名寄せの結果つけられた機関 ID を比較することによって行った.

また，日本人の名前を英字表記にした場合，ローマ字表記として，ヘボン式，ローマ式，訓令式など複数存在するため，英字表記の種類は多くなる. さらに，データの中には，姓がヘボン式で名が訓令式などの変形パターンも存在した. そのため，研究者 1 人ずつに対して，これら複数のパターンを組み合わせたデータを準備する必要があった.

(3) マッチング処理

アルファベット表記の研究者名のみでの名寄せ処理は簡単である. しかし，例えば，「さとうひろし」といった，日本人に多い名前の場合は，名前だけで同一人物と判定することはできない. そのため，本プロジェクトにおいては，姓と名の一致の他に所属機関の一致を条件とした.

研究者と所属機関の関連付けについては，論文情報と特許情報ではデータの性質上異なる手法を用いた. 論文情報では，研究者とその所属機関はリンクした形で文献データベースに格納されていた. しかし，特許データベースでは前述のとおり機関も研究者も同列の扱いになっている. 研究者の名寄せについては，所属機関の情報が必要なので，以下のように処理を行うこととした.

それぞれの特許において，出願人情報に機関名が，発明者情報に研究者名が含まれていると考える. ある研究者の特許を調べる場合には，発明者情報の中に，その研究者の「姓」「名」の両方のデータが含まれ，その特許の出願人情報が，その研究者の所属機関と一致している場合に，その特許の発明人をその研究者と判定する. あるいは，研究者名の一部として機関名が含まれている場合（例えば，「HOSONO, Hideo, c/o Tokyo Institute of Technology」）場合には，その機関名も利用する. しかし，研究者の所属機関が必ずしも出願人となっているとは限らないため，この方法ではまだ改善の余地があると考えられる.

また，今回の方法は，研究者の所属機関が分かっていることを前提としている. あらかじめ名寄せ対象者名を絞った本プロジェクトでは，そのデータの入手が可能であったが，この方法では全ての研究者の名寄せを行うことはできない. また，所属機関を一致の条件として使う場合，精度を上げるためにはその所属機関に所属していた期間を考える必要がある. しかしながら，研究者の所属機関について正確に記述された履歴情報は存在せず，整備には人手も時間も膨大にかかるため，今回は所属していた期間については考慮しないことにした.

研究者の名寄せの精度を上げるためには，その他に共同研究者や活躍する学術分野などについても考慮して処理する必要があり，今後の課題となっている.

5. 特許の引用情報にみられる非特許文献情報と論文情報のマッチング

5.1 非特許文献情報

特許データベース内の非特許文献情報は単純な文字列データとして 1 フィールドに格納されており，文献データベースに存在する論文とマッチングをかけるために必要な，タイトル，雑誌名，人名，ページ数などの項目に区切られていない. そのためマッチングに際してこれらの項目を抜き出す必要があった. しかし，各項目にあたるデータに様々なパターンが存在するだけでなく，項目の出現順序にも様々なパターンがあったため，項目を特定しマッチングを行うことは非常に困難な作業であった. 例えばページについては，論文情報では表 4 のような表記パターンが，非特許文献情報では表 5 のような表記パターンがあり，それぞれに対応する形でマッチングをかける必要があった.

表 4　ページ情報の表記パターン例（文献データベース）

S86	II
S406	X
L119	S-C5
x6	vii
G969	FT11
viiS	S-87
P.44	xv
XC	ss8
i1	S-21
PL187	I-5
NIL.0001	UNAIDS1

表 5　ページ情報の表記パターン例（特許データベース）

pages L 473 - L 475	page 649 - page 651
pages S885 - S890	page 665-668,670-672
pages 339 - 343	page 15
pages 37-38,40,42	PP.281-3
pages 99-100,102	PP.41, 56
pages 145-148	pp. (69)21 - (75)27
pages A,22-36	pp. 223-224 and 235-236
pages 5149-52	pp. 685 and 1039
pages 28, 30-31	pp.82, 83, 112
pages 656 to 658	p. 12-15
pages 083114-1 - 083114-5	P-720

また，非特許文献情報自体にもデータの一部が欠損していると思われるもの（表 6 参照）などが存在した．この問題については本プロジェクトでは根本的な対策ができず，改善については次期プロジェクトでの課題となった．

表 6　データ欠損のある非特許文献情報の表記例

表記	説明
No further relevant documents disclosed	引用情報が伏せられている
US-A-16 945 993	雑誌などの情報が見あたらない
None	引用情報が存在しない
J. HETEROCYCLIC. CHEM., Bd. 4, Nr. 4, 1967, Seite 565-567 XP002015652 FEHNEL: Friedl{nder Synthesis with ...	途中で引用情報が切れている

5.2　マッチング処理

文献データベースと，特許データベース中の非特許文献情報をマッチングするために，両者に共通した項目の探索を行った．メールアドレスは個人を特定する情報として精度が高いが，どちらのデータベースにも記入されている割合が非常に低く項目から除外した．また人名（人名，発明者）や機関名も有力な項目であるが，辞書内の表記パターンが多く，膨大なデータのマッチングには時間がかかりすぎるという理由で除外した．検討を行った結果，雑誌名，開始ページ，終了ページ，発行年，巻号情報を主として使用することにした．しかし，マッチングに使用する項目としてすべて揃っていないケースも多く存在したため，その際は補助的に人名を使用し，できる限り多くの項目を使ってマッチングが行えるよう配慮した．

また，マッチングの前処理として，機関名寄せの際に使用した略記辞書を使った変換処理を行い，マッチング処理の精度が高くなるようにした．

6.　論文情報および特許情報抽出用検索システムの構築

論文情報および特許情報の人名および機関名についての名寄せを実施した結果について，データベースを整理し，論文・特許情報をキーワードや人名などで抽出できるように検索システムを構築した．

なお，検索対象となる論文情報については，それぞれ被引用数と被引用数のパーセンタイルをデータとして付加した．被引用数のパーセンタイルは，ASJC† 毎に集計された論文を被引用数の多い順に並べた場合に，その論文が上位何パーセントに位置するかを示したもので，数値が低いほど被引用数が多いといえる．また，論文情報については特許情報からの引用数についても被引用数と被引用数のパーセンタイルを付加した．これらの数値は，特許情報からどのような論文情報が参照されているかを推測するための一指標として用いることを目的に導入した．

† 文献データベースにて，文献情報の学術区分を表す区分コード．

図 1 は論文情報および特許情報抽出システムの入力画面である．この画面内にてデータ抽出を行いたい論文情報または特許情報の検索項目を入力し，「検索実行」ボタンを押すと，検索が実行され論文情報または特許情報を別画面にてダウンロードできる（図 2）．人名または機関名については，名寄せ対象であったものに限り検索項目として指定できる．

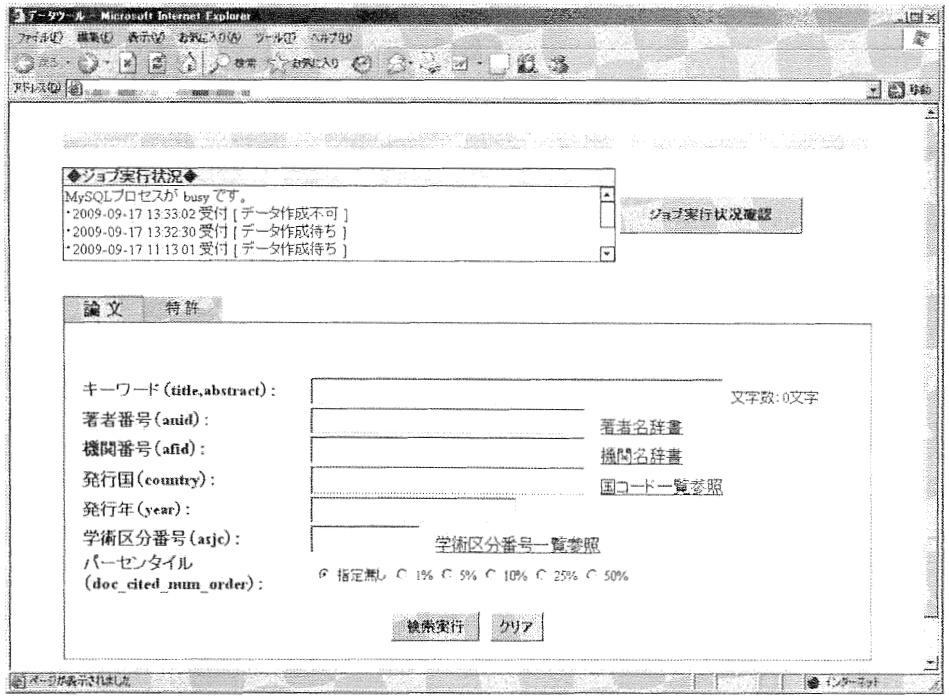

図 1　論文情報および特許情報抽出システム（入力画面）

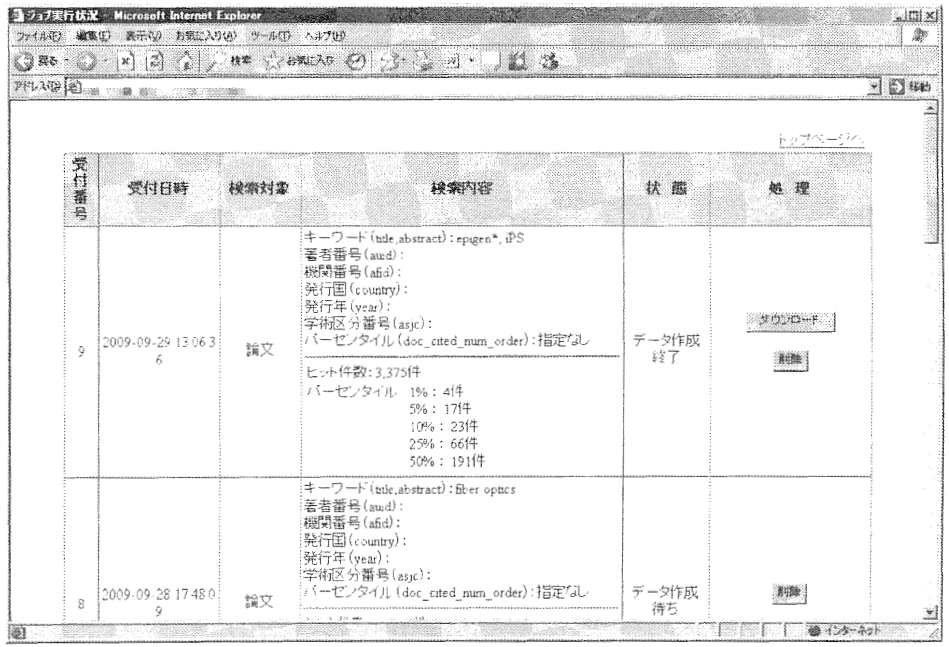

図 2　論文情報および特許情報抽出システム（検索結果画面）

さらに，ダウンロードデータには論文と特許の内容の可視化分析を行うため，可視化ソフトウェアと連動するためのデータも追加した．ダウンロードデータにある描画用データを NRI サイバーパテント社の「TRUETELLER パテントポートフォリオ」に読み込ませることで，図 3 のような図を描くことができ，特許の引用情報にみられる論文情報を視覚的にも分析することが可能になるようにした．

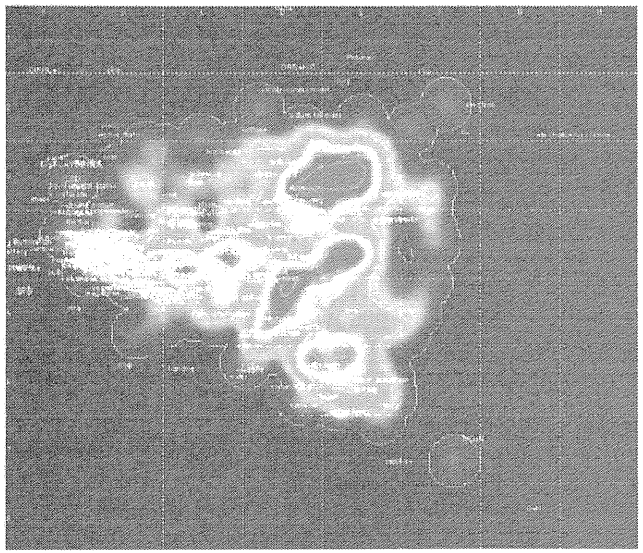

図 3　可視化例

7. おわりに

現在，特許情報に引用される論文情報の特徴については調査分析を行っている段階だが，学術分野や研究機関の特性がある程度みられるようになってきている．文献データベースおよび特許データベース内のデータについての名寄せについては，さらに精度を向上させる方法はあるものの，実現のためには処理に必要な開発環境を整備する必要がある．今後，このプロジェクトで得た知見を元により適切な開発環境を整備して精度の高いデータベース作成を目指していきたい．

参考文献

1)　岡崎輝男: 特許データベースの課題, 特技懇, Vol.250, pp. 97-105 (2008).