Automatic Translation of Scholarly Terms into Patent Terms

HIDETSUGU NANBA,†1 HIDEAKI KAMAYA,†2 TOSHIYUKI TAKEZAWA,†1 MANABU OKUMURA,†3 AKIHIRO SHINMORI†4 and HIDEKAZU TANIGAWA†5

In this paper, we propose a method to translate scholarly terms into patent terms (e.g., translating “word processor” into “document editing device” or “document writing support system”). The method is useful when users search both research papers and patents in a particular field. To translate scholarly terms into patent terms, we proposed two methods: “citation-based method” and “thesaurus-based method”. We also proposed a method combining the two with an existing “Mase’s method”. To confirm the effectiveness of our methods, we conducted some examinations, and found that the combination method performed the best in terms of Recall, Precision, and ϵ, which is an extensional measure of MRR (Mean Reciprocal Rank) widely used for the evaluation of question answering.

1. はじめに

本稿では、論文で使われる用語（以後、論文用語1）を特許中で使われる用語（以後、特許用語）に自動変換する手法を提案する。これは、たとえば「ワードプロセッサ」という論文用語を入力すると、「文書編集装置」や「文書作成支援装置」といった特許用語に自動的に変換する技術2のことであり、ある分野の特許と論文を同時に検索しなければならない状況でこのような技術を使えば、ユーザーの検索作業を支援することが可能になる。

近年、知的財産権に対する関心が高まり、企業はもちろん、個人が特許を取得するケースも増加している。特許出願の際には、出願しようとする発明に新規性があるかどうか、他者の権利を侵害していないかといった調査、いわゆる「先行技術調査」が必要不可欠である。先行技術調査は、出願された技術が特許権の取得に該当するかどうかの判断をするために、特許庁の審査官がきわめて厳密に行う実体審査である。この実体審査では、過去に同様の出願技術が存在していたかどうかが徹底的に調査される。企業においては、競合する他者の権利を無効化するために、特許庁の審査官による審査を経た出願技術を再調査するといった社内調査が行われている。このような社内調査は無効資料調査と呼ばれる。こうした調査では、特許や論文など様々なジャンルの文献が検索の対象になり、通常は、サーチャーや検索エンジンの検索に限らず、個々の特許の専門家がこの調査の担当をしている。しかしながら、審査官やサーチャーは、文献のジャンルに応じて、適宜表現を使い分けて検索を行う必要があ
論文用語の特許用語への自動変換

本稿では、提案する技術が利用される場面として、このような論文と特許を検索対象とした先行技術調査や無効資料調査を想定している。

論文と特許を対象とした検索は、特許庁の審査官や企業などのサーチを限ったことではない。近年では、大学などの研究者にとって、特許出願が研究活動の一つとして重視されるようになってきており、研究者が特許と論文を検索する機会が増えてきている。しかし、特許では権利範囲をなるべく広く確保するため、一般性の高い特許用語を用いて記述する傾向がある。このため、単純に表層的な単語の一一致を見なすだけである従来の検索モデルでは、同じキーワードで特許データベースと論文データベースを検索しても、用語の使いわけ方の違い、そのキーワードに関する論文や特許を十分に収集できるとは限らない。論文用語の特許用語への変換技術は、このような場合においても有効であると思われる。

論文用語を特許用語に変換する技術は、のほかにも特許作成の場面でも利用可能である。上述のとおり、特許中では一般的な用語が使われる傾向にあるが、特許を作成する際に、特許中で用いる用語をどのように一般化するのか、について検討する必要がある。同じ論文用語でも、どの点に着目するかにより、用いる特許用語が異なる場合がある。たとえば、特許中で「フロッピーディスク」について言及する場合、「持ち運び可能である」という特徴に着目する場合には「リムーバブル記録メディア」、「磁気で記録する」という特徴に着目する場合には「磁気記録メディア」といった用語が用いられる。このような場合において、論文用語の特許用語への変換技術を用いることで、特許を作成する作業を支援できる。

本稿の構成は以下のとおりである。次章では、本研究で提案する手法が実際にどのような場面で利用されるのかを示すため、著者らが開発中の特許・論文検索システムを示し、その動作例を説明する。3章では関連研究について述べる。4章では、論文用語の特許用語への変換手法を提案する。5章では、提案手法の有効性を調べたために行った実験について述べる。最後に6章で本稿をまとめめる。

2. システム動作例

本章では、論文用語を特許用語に自動変換する機能を取り入れた検索システムについて、その動作例を紹介する。図1 は、キーワード検索の画面であり、画面の左側が論文検索用、右側が特許検索用のフォームになっている。以下では、特許検索に依頼しないユーザが特許を検索したり、特許業務に携わる専門家が誤って検索しない分野の特許を検索したりする場合の一般的な操作手順について説明する。まず、画面左側の論文検索フォーム内の「タイトル」欄（図1 (1)）に、キーワードを入力する（図1 の場合「フロッピーディスク」という論文用語が入力されている）。この状態で「関連特許用語」ボタン（図1 (2)）を押すと、「関連する特許用語のリストがポップアップウィンドウ中に表示される（図1 (3)）。図1 では、「フロッピーディスク」に関連する特許用語として「記憶媒体」、「記憶媒体」、「ディスク記録メディア」、「といった用語が表示されており、ユーザがこれらの用語を選択すると、その用語が特許検索フォーム内の「発明の名称」の欄（図1 (4)）に追加される仕組みになっている。

特許検査に反対のないユーザでも、検索用キーワードとして提示された候補の用語が適切であるかどうかはある程度判断できる。「記憶媒体」、「ディスク状記録メディア」といった用語そのものが思い浮かばなくても、本システムを利用すれば、「フロッピーディスク」という論文用語を入手するだけで、関連する特許用語の候補を得られるため、「フロッピーディスク」を含まない関連特許を検索することが可能となる。本稿では、この論文用語を特許用語に自動変換する手法を提案し、実験によりその有効性を検証する。

3. 関連研究

本章では、「ジャンル横断情報アクセス」と「特許と論文を対象にした分析・調査」に関する関連研究について述べる。

3.1 ジャンル横断情報アクセス

ジャンル横断情報アクセスに関して、これまでに様々な研究が行われているが、その1つに第3回NTCIRワークショップで実施された技術動向調査タスクがあげられる7)。このタスクでは、入力された新聞記事に関連する特許を特許データベースの中から検索する。ここで、以下の点を考慮することがポイントの1つとなる。たとえば「社長」という単語は新聞記事中には比較的高頻度で出現するが、特許中にはほとんど出現しない。このため、情報検索分野において一般的に使われる逆文書頻度（IDF）を単語の単位付け手法として用いる場合、またも特許に「社長」という単語が出現すると、非常に高い重みを与えてしまう、という問題が生じる。

このタスクにおいて、Itoらは、「Term Distillation」という概念に基づいた新しい検索手法を提案している8)。「Term Distillation」とは、上述の「社長」という単語の例のようなジャンルによる単語の出現頻度分布の違いを、単語の重み付けの際に考慮することで、より

1 このプロトタイプシステムでは「発明の名称」の欄に特許用語が追加されるが、「請求項」や「断面図」や「要約」の欄に特許用語を追加するように変更することも可能である。
Fig. 1 System snapshot.
リ精度の高いジャンル横断情報検索を実現する手法のことであり、この考え方は、新聞記事とブログ記事の対応付けに有効であることがIkeedaらにより確認されている\(^1\)。しかし、この方法は、たとえば「磁気記録媒体」のように特許中では一般的に使われるが論文中ではまったく使われない用語に対しては適用できない。

Nanbaらは、本稿で提案する用語の変換とは別の側面からジャンルを横断した情報アクセス技術の開発に取り組んでいる\(^2\)。近年、特許中で関連論文を、逆に論文において関連特許を、引用するケースが増えているが、このような文書間の引用関係をたどれば、論文と特許を横断して関連文書を収集することができる。Nanbaらは特許中で引用されている論文の書誌情報を抽出し、それを論文データベース中の書誌情報と比較・同定することにより、特許と論文データベースを統合する手法を提案している。ただし、現状では、特許中の引用文献の中で特許が占める割合は数パーセント程度であるため、あるテーマで特許と論文を網羅的に収集するのには、特許と論文間の引用関係をたどるだけでは限界がある。そこで、本研究では論文用語の特許用語への変換に取り組む。

特許と論文を対象にしたジャンル横断情報アクセスに関するこのほかの研究プロジェクトとして、国立情報学研究所主催の第7回NTCIRワークショップで実施されている特許マイニングタスクがあげられる\(^3\)。このタスクでは、日本語または英語論文抄録を、特許分類体系の1つである「国際特許分類」(International Patent Classification: IPC)に自動分類することを目的とする。特許を分類するタスクは、これまでに第5回および第6回NTCIRワークショップにおいて、Fターム分類タスクとして実施されてきたが、特許マイニングタスクでは、分類対象となる文書が論文に変わるため、特許と論文で使われる用語の違いについて新たに検討する必要がある。現在のタスクは実施中であり、2008年12月の成果報告会で、各参加グループの技術詳細が報告される予定である。本研究で目指している論文用語と特許用語への変換技術は、このようなタスクにも応用することが可能である。

3.2 特許と論文を対象にした分析・調査

特許と論文を対象にした分析・調査の1つに、1章で述べた先行技術調査や無効資料調査がある。無効資料調査の自動化の試みとして、これまでに、第4回\(^4\)，第5回\(^5\)，および第6回\(^6\)NTCIRワークショップにおいて、無効資料調査タスクが実施されている。第6回ワークショップでは、5つの団体が日本語の公開特許公報を対象としたタスクに参加し、21システムを提出している。これらのシステムはMean Average Precision（MAP）により評価され、最も良いシステムでMAP値0.0815を得ている\(^7\)。このシステムは、入力となる請求項（検索クエリ）の構造を解析し、特定の個所の用語に対し、IDFなどの重みを与えることで、上記の精度を達成している。

これまでにNTCIRで実施してきた無効資料調査タスクでは、検索対象が特許文書に限定されており、論文は含まれていない。一方で、前章でも述べたとおり、実際に行われる無効資料調査では、特許だけでなく、論文も調査の対象となるため、論文も対象としたテストコレクションの構築が必要とされている。

近年、特許と論文を対象にした検索のニーズが増えるつつあること、また、2006年から毎年、政府の知的財産戦略本部が発表している「知的財産推進計画」において、大学研究における特許情報の重要性が課題となっていることも指摘し、特許と論文を検索できるシステムの開発やサービスの提供がいくつか始まっている。Tomson・サイエンティフィック社のISI CrossSearchでは、様々な分野の学術雑誌、国際会議の予稿集、世界40か国の特許発明機関から収集した特許データベースなどを検索することができる。富士ゼロックスのDocuPatでは、日米特許データ1,800万件と科学技術振興機構（JST）が提供する科学技术文献データ2,000万件を1つのインターフェースで検索することが可能である。JSTでも、特許と論文が検索可能なシステムJSTPatMを開発し、2007年3月から大学向けにサービスを開始している\(^8\)。NRIサイバーパートナーズのTRUE TELLERは、マッピング分析やグルーピング分析など様々な観点から特許を分析することができるが、CSV形式の論文データを取り込むことにより、論文と特許を対象にした分析も可能となる。

しかしながら、これらのサービスでは特許と論文用語の変換機能は、少なくともこれまで提供されていない。このため、あるテーマに関する特許と論文を網羅的に収集するには、ユーザ自身が特許と論文用語の違いの問題を解決する必要がある。こうした既存の検索システムの改善に、本研究技術が利用できると思われる。

4. 論文用語の特許用語への自動変換

本研究では、「引用手法」、「シーソラス手法」、「Maseらの手法」という3種類の手法を組み合わせた論文用語の特許用語への変換手法を提案する。4.1節で引用手法、4.2節でシーソラス手法、4.3節でMaseらの手法について、それぞれ述べる。また、4.4節でこれらの3手法を組み合わせた用語変換方法について述べる。

\(^1\) http://pr.jst.go.jp/patmreg
4.1 特許、論文間の引用関係を用いた用語変換

難波[2]は、ある用語を入力すると、それに関連する用語を自動収集する方法を提案している。この手法では、まず、ある用語を表題に含む論文を収集し、次に、それらと直接引用
関係にある論文の表題から用語を抽出し、最後に、それらを頻度順に並べて出力している。同様に、ある用語を表題に含む論文を収集し、それらと直接引用関係にある特許から特許
のトピックを示す用語を抽出すれば、入力された論文文脈に関する特許用語の変換が実現
できると考えられる。

本研究では、Nanba らの手法[14]で得られた特許、論文間の引用関係データを用い、以下の手順で、論文用語を特許用語に変換する。なお、この手法を、以後本稿では引用手法と呼ぶことにする。

(1) システムに論文用語を入力。
(2) システムは、入力された用語を表題に含む論文データベースから検索。
(3) 手順 (2) で検索された論文と引用関係にある特許を収集。
(4) 手順 (3) で収集された特許から用語を抽出し、頻度順にならべ、出力。

ここで、手順 (4) において、特許中のどの箇所から用語を抽出するのかも検討する必要
がある。本研究では、特許から用語を抽出する際、著者名に着目し、特許东南亚を得る者の、特許
を受けようとする発明を特定するために、必要と認める事項のすべてを記述した「特許文」と、特許
明細書の中心で最も重要な箇所である、また、権利範囲をなるべく広く確保する
ため、請求項では一般的な特許用語を用いて記述されるという特徴がある。

図 2 は請求項の一例であるが、この例から請求項内には数多くの用語が存在し、そのす
べてを抽出すると関連のない用語も抽出結果に数多く含まれてしまうことが分かる。ここ
で、請求項には以下に述べるような 2 つの構造的な特徴が存在する[16]。1 つ目は、請求項
の記述末尾に数形形解析器 ChSen で名詞または記号という品詞が付与された形態素が存在
し、その直前に名詞があり、さらにその直前に、名詞、記号、または助詞「の」が連続的
に出現して名詞句を形成する、という特徴である。2 つ目の「において」や「であって」など
の文字列を用いて記述を前半と後半に分割するとき、「において」や「であって」
の直前に、記述末尾と同様の名詞句が存在する、という特徴である。このまわりや、発
明の明確名を表していることが多い。新垣らは、手がかり論文を利用して請求項の構造を解析する
手法を提案しているが、この解釈手法を用い、図 2 の文に該当する箇所を特許用語とし
て抽出する。

本研究では、このほか、特許申出時請求項文の関係において着目する、特許申出時、複数の独立
請求項（他の請求項を引用しない請求項）を含む独立請求項を引用する従属請求項が存在
する。複数存在する独立請求項の中でも特に、第 1 請求項で、その主申出時における最も
大きな発明について述べる傾向にあることから、用語抽出の対象となる請求項を、第 1 請求
項（独立請求項）を選択する必要がある。特許申出時文は、一般的な特許
用語を用いて記述されるため、特許申出時文の理解と、特許申出時文の理解
を利用した、特許申出時文の理解と、特許申出時文の理解など、従属請求項を利用し、第 1 請求項を
使用場合、特許申出時文の中のすべての請求項を使う場合の 3 通りが存在するが、今回は、
予備実験の結果から、入力された用語を含む論文と引用関係にある特許の第 1 請求項（独立
請求項）とその従属請求項文を用いる)。

4.2 請求項の上位/下位関係を考慮した用語変換

4.1 節でも述べたように、特許では、権利範囲をなるべく広く確保するため、しばしば一般性の高い特許用語が使われる。つまり、特許用語の多くは論文用語の上位用語であると考え
られる。そこで引用手法は別に、特許シノラスを用いた上位用語の収集による手法を
提案すると。このシノラスには、Nanba が自動的に構築したものを利用する[13]。Nanba は
「A や B などの C」といった定型表現[1]に着目し、公開公報（1993～2002 年）から、これ
らの表現を含む文を抽出し、異なり数 1,825,518 言、上位/下位関係 7,031,159 から構成
されるシノラスを構築している。このシノラスには、公開公報の中での各上位/下位関
係の出現頻度が付与されている。
以下の手順で上位/下位シナロアを用いた特許用語の収集を行う。なお，この手法をシナロア手法と呼ぶ。

(1) システムに論文用語を入力，
(2) システムは，人力された用語の上位語を上位/下位シナロアから収集，
(3) 手順 (2) で得られた用語セットを頻度順に並べ，出力。

4.3 Mase からの手法を用いた用語変換

特許明細書の「符号の説明」という項目は，特許中で例示されているシステムなどの構成要素について説明する項目であり，たとえば「磁気記憶装置（フロッピーディスク）」といった記述が存在する。Mase からの解析例を引用する。

以下に，Mase からの手法による引用手法とシナロア手法の比較手法について述べる。

1 引用手法，シナロア手法いずれも用語の出現頻度を各用語のスコアとして，スコアの大きいものから順に出力されるが，それぞれの用語のスコアを，順位 1 位のスコアで割ることにより，順位が 1 位の用語のスコアが 1 となるように正規化する。

2 Mase からの手法により抽出された各用語の，「符号の説明」の項目内での頻度および用語の末尾の名詞（語尾）を取り出す。

3 手順 2 で抽出された末尾名詞ごとに頻度を合計したうえで，上位 1 位のスコアが 1 となるように正規化する。

4 引用手法またはシナロア手法で得られた各用語の末尾の名詞が手順 3 の末尾名詞と一致する場合，手順 2 で得られた引用手法とシナロア手法による各用語のスコア，手順 3 で得られた名詞の値を加算し，引用手法とシナロア手法の結果をスコアが大きい順に並べる。

なお，手順 4 の前提では，Mase からの手法による末尾名詞のスコアと引用手法およびシナロア手法のスコアをそのまま加算しているが，加算の際に Mase からの手法のスコアに重みを与えようという目的により，Mase からの手法の引用手法やシナロア手法への影響の度合いを変えることができる。この重みの決定方法については，5.1 節で述べる。

引用手法とシナロア手法の組合せ

Mase からの手法により改良された引用手法とシナロア手法を組み合わせた方法（以後，統合手法）について述べる。ある論文用語を引用手法とシナロア手法で特許用語に変換し，
5. 実験

提案手法の有効性を確認するため実験を行った。本章では、5.1 節で実験条件について説明し、5.2 節で実験結果を報告し、5.3 節で結果を考察する。

5.1 実験条件

実験データ

実験には公開特許公報（1993－2002年）を用いた。また、特許、論文間の引用関係の情報は、Nanbaからの手法(4)を用いて抽出した特許中での引用論文の書誌情報約85,000件を用いた。

正解データは以下の手順で作成した。

(1) 特許中で引用されている論文の著者情報85,000件中から名詞を抽出し、頻度順に並べる。
(2) その中から論文用語を人手で選択する。
(3) 論文用語ごとに、提案手法および後述するベースライン手法を用いて特許用語の候補を抽出する。
(4) 手順（3）で得られたすべての候補の中から正解用語を人手で判定する。

なお、手順（4）では、以下に示す4つの基準に基づいて正解判定を行った。

・論文用語と「全体・部分」関係にある特許用語は不正解とする。たとえば、論文用語「ワードプロセッサ」の場合、「文書編集装置」は正解であるが、その構成要素である「表示装置」は不正解と判定する。

・ある特許用語の候補の特許データベース中の出現頻度が極端に低い場合には、その用語は特許において一般的な表現ではなく、特許検索を行ううえで有用でないと考え、不正解とする。

・表1 評価用データの例

| 論文用語（人手） | 特許用語（出力） |
|-----------------|------------------|
| DRAM            | 半導体メモリ．半導体メモリ装置． |
| メモリセル      | 半導体メモリ装置． |
| 光磁気ディスク  | ディスク状記録媒体．円盤状記録媒体． |
| ワードプロセッサ | 文書処理装置．文書情報処理装置． |
| ノボラック樹脂   | オルガナラセンブリ． |
| テレビカメラ     | 影像情報装置． |

*1 http://kantan.nexp.jp/
論文用語の特許用語への自動変換

表 2 評価対象の分類

| IPC | 頻度 | IPC の説明 |
|-----|------|------------|
| H01L | 8 | 半導体装置または固体装置など |
| C12N | 7 | 微生物または酵素 |
| G11B | 7 | 記録ラベルと変換器の間の相対の運動に基づいた情報記録 |
| G06F | 6 | 電気的デジタルデータ処理 |
| A63F | 2 | 記憶のための特別の適用に特徴があるデータ処理装置 |
| G09G | 2 | 静止的手元を通じて可変情報を表示する表示装置の制御のための装置または回路 |
| G02B | 2 | 光学装置、光学系、または光学装置 |
| B41M | 1 | 印刷、複製、マーキングまたは複写方法 |
| A61B | 1 | 診断・手術、個人識別 |
| G03F | 1 | フォトメカニカル法による凸凹化またはパターン化された表面の製造 |
| H01M | 1 | 電池 |
| C09D | 1 | コーティング組成物 |
| H05K | 1 | 印刷関係 |
| G01N | 1 | 材料の化学的または物理的性質の決定に関する材料の調査または分析 |
| G06Q | 1 | 管理目的、商用目的、金融目的などに特に適したデータ処理システムまたは方法 |
| H04N | 1 | 図書通信 |
| H03H | 1 | インテリジェントな通信系 |
| H01S | 1 | 読出放送を用いた装置 |
| G03B | 1 | 写真を撮影するためのまたは写真を投影もしくは反射するための装置または配置 |
| B41J | 1 | タイプライタ、選択的プリンティング機器 |

と、その中でさらに学术と関連する分野は限られていることを考慮すると、表 2 に示す程度の偏りは、許容範囲内であると考えられる。

評価尺度

評価には、以下に定義される 

\[
\epsilon = \frac{1}{\sum_{i \in R} \frac{1}{j}}
\]

この定義は、質問応答システムの評価で一般的に用いられる MRR (Mean Reciprocal Rank) を拡張したものであり、システムが出力結果の上位に数多くの正解を出すければ、1 に近い評価値が得られる。

提案手法

(1) 引用手法 (Cite)
(2) 1+Mase らの手法を用いた (1) の改良手法 (Cite(M))
(3) ソーラス手法 (TheS)
(4) Mase らの手法を用いた (3) の改良手法 (Thes(M))
(5) (2)+(4) の統合手法 (Cite(M)+Thes(M))

ベースライン手法

(6) Mase らの手法を単独で用いた手法 (Mase)
(7) 入力された用語と関連する用語を抽出 (GETA)
(8) 同義語抽出手法で用語を抽出 (Syn)
(9) JST ソーラスを用いて用語を抽出 (JST)

手法 (1)，(3)，(5)，(6) は，それぞれ 4.1 節，4.2 節，4.3 節，4.4 節で説明した手法である。また，手法 (2) と (4) は，それぞれ手法 (1) と (3) を，4.3 節で述べた Mase らの手法で改良したものである。なお，手法 (2)，(4)，(5) で用いるパラメータの調整方法については後述する。

ある用語 A と高頻度で共起する用語は，用語 A と関連度が高い，という考えに基づき，ベースライン手法の 1 つとして，手法 (7) では入力された用語と共起頻度の高い用語を出力
論文用語の特許用語への自動変換

する。なお、共起語の抽出には汎用連想検索エンジン GETA*1を利用する。

2つ目のベースライン手法として、自動的に構築された同義語辞書*3を用いる。一般に、文書中では同義語が自発表現で示されることが多い。たとえば、「ワードプロセッサ（ウィ-
プロ）」といった表現の場合、この同義語として「ワードプロセッサ」の同義語として「ワード-
プロ」が抽出される。しかし、自発的用法は多様であるため、「（ ）」という形式の同義語を
すべて抽出すると、数多くのノイズが含まれてしまう。この問題を解決するために、4.3節で
述べた Mase らの手法では、抽出個所を特許内のか（付録の説明」という項に限定することで
、高い精度に同義語抽出を実現している。本研究では、提案手法にあたる Mase らの手法
とは別のベースライン手法の 1 つとして Nanba の手法を用いる*3。4.2節で述べたと
おり、Nanba は、特許中の「A」や「B」などの「C」といった定型表現に着目し、上位/下位シ-
ソラスを構築しているが、このような表現の中で、たとえば「ハードディスク（磁気記憶
装置）」や「光ディスク」の記憶装置」といった表記があった場合、Nanba は「ハードディス-
ク」と「磁気記憶装置」を同義語として抽出している。また、こうして自動的に構築され
た 50,161 個の用語対を特許検索の検索質問拡張に利用することで、検索精度が向上することを実験により確認している。本研究では、この辞書をベースライン手法 (8) として用いる。

3つ目のベースライン手法として、JST が作成し、オンライン上での一部（約 40,000
語）を公開しているシンソラス (1999 年版)*2を利用して*3。このシンソラスを用いることで、
ある論文用語の上位/下位語や同義語を調べることができる。本研究では、シンソラスの上
位/下位語に対し、4.2節で述べたシンソラス手法と同様の手順で、ベースライン手法 (9) と
して、特許用語への変換を行う。

提案手法 (2), (4), (5) におけるパラメータ

提案手法 (2) と (4) について、Mase らの手法による重みを、引用手法およびシンソラス
手法にそれぞれどの程度与えるべきかを決めるため、予備実験を行った。予備実験には、上
述の評価用データとは別に用意した論文用語 25 語（および対応する特許用語）を用いた。
$m = 0$ より 1 まで 0.1 刻みで変えて、引用手法とシンソラス手法それぞれについて、$\epsilon$ を調
べた結果、提案手法 (2) では $m = 0.8$, (4) では $m = 0.2$ のときに評価値 $\epsilon$ が最大となった。

1 http://geta.ex.nii.ac.jp/
2 http://jdrdream2.jst.go.jp/html/thesaurus99/thesaurus_index99.htm
3 JST シソラスは、元は論文用に作成された概念体系であるが、上位概念の用語になるほど、一般に実際の論文
中の使用頻度は増し、一方で、特許中ではるべく権利範囲を広げるため、上位概念の用語が使われる傾向に
ある。このため、JST シソラスは、上位概念の用語に関しては、特許検索用にも利用できると考えられる。

| 提案手法 | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Cite(M)  | 0.136  | 0.173  | 0.231  | 0.240  | 0.298  | 0.107  | 0.011  | 0.058  | 0.050  |

| 提案手法 | Cite(M) | Thes(M) | Cite(M)+Thes(M) | Mase | GETA | JST |
|----------|---------|----------|-----------------|------|------|-----|
| (2)      | $\epsilon$ | 0.151  | 0.165  | 0.170  | 0.173  |
| (4)      | $\epsilon$ | 0.109  | 0.242  | 0.275  | 0.311  |
| (6)      | $\epsilon$ | 0.115  | 0.073  | 0.056  | 0.047  |
| (8)      | $\epsilon$ | 0.213  | 0.235  | 0.229  | 0.240  |

| 提案手法 | Cite(M) | Thes(M) | Cite(M)+Thes(M) | Mase | GETA | JST |
|----------|---------|----------|-----------------|------|------|-----|
| (2)      | $\epsilon$ | 0.151  | 0.165  | 0.170  | 0.173  |
| (4)      | $\epsilon$ | 0.109  | 0.242  | 0.275  | 0.311  |
| (6)      | $\epsilon$ | 0.115  | 0.073  | 0.056  | 0.047  |
| (8)      | $\epsilon$ | 0.213  | 0.235  | 0.229  | 0.240  |

次に、提案手法 (5) における引用手法と Mase らの手法の統合方法について述べる。上
記の予備実験用データを用い、4.4 篇「引用手法とシンソラス手法の組合せ」で述べたパラメー-
ータの値を 0 から 1 まで 0.1 ずつ増やし、各段階において、$\epsilon$ 値を算出した。予備実験
結果、$\lambda = 0.3$ のときに $\epsilon$ 値が最も高かったので、以後はこの値を用いて実験を行った。

5.2 実験結果

各手法による $\epsilon$ による評価結果を表 3 に示す。また、提案手法で結果の良かった (2), (4),
(5), ベースライン手法で結果の良かった手法 (6), (8) について、さらに詳細に評価した。
表 4 に結果を示す。表では、提案手法およびベースライン手法とともに、システムが理想的
な出力を行った場合の $\epsilon$ の値を、それぞれ示す。また、表 4 で、$\epsilon$ 値とあわせて、シス템
表 5  引用手法 (2), ショーナス手法 (4), 結合手法 (5) の比較
Table 5  Comparison of system outputs by the citation-based method(2), the thesaurus-based method(4), and the combination method(5).

|                | 再現率 | 精度 | \( \epsilon \) |
|----------------|--------|------|---------------|
| \( C+T(5) < \text{Thes}(4) \) | 2      | 2    | 14            |
| \( C+T(5) > \text{Thes}(4) \) | 13     | 13   | 17            |
| \( C+T(5) = \text{Thes}(4) \)  | 31     | 31   | 16            |
| \( C+T(5) < \text{Cite}(2) \) | 3      | 4    | 9             |
| \( C+T(5) > \text{Cite}(2) \) | 23     | 23   | 25            |
| \( C+T(5) = \text{Cite}(2) \) | 21     | 20   | 13            |

表 6 手法ごとに上位 20 件に正解が 1 件も抽出されない事例の割合
Table 6  The ratio of cases that each system could not correctly convert scholarly terms within top 20.

|                | Cite(M) | Thes(M) | Cite(M)+Thes(M) | Maese | Syn |
|----------------|---------|---------|-----------------|-------|-----|
| \( \text{Cite}(2) \) | 48.9% (21/47) | 40.4% (19/47) | 25.5% (12/47) | 55.3% (26/47) | 74.5% (35/47) |

法 (2) を統合することにより改善してしまった事例は全体 (全 47 件) のわずか 4.3% (2 件) であり, 逆に再現率と精度が向上している事例が 27.6% (13 件) なみで, 同様の調査を引用手法(2) と結合手法(5) 間でも行ったところ, 再現率と精度では, 結合手法(5) が引用手法(2) およびショーナス手法(4) を大幅に改良できていることが分かった (表 5). 以上の結果より, 今回提案した手法(5) の統合方法が妥当であったと結論付けることができる. 各手法の再現率
全体の約3/4の場合においてこの手法がユーザの役に立つことを意味している。統合手法（5）におけるこの失敗事例12件のうち、半数は化学系の用語*1であった。件数が少ないので、この結果だけでは断定することはできないが、分野によって本提案手法の変換精度に、無視できない程度の差が生じている可能性が示唆される。

6. おわりに

本稿では、論文用語を特許用語に自動的に変換する手法として、特許と論文間の引用関係を利用した「引用手法」と自動的に構築した上位/下位リンクソースを利用した「シノリャス手法」、およびこれらの2手法と既存手法のMaseらの手法を組み合わせた用語変換手法を提案した。提案手法の有効性を調べるため実験を行った結果、質問応答システムなどの評価で一般的に用いられるMMRを用いた尺度でおよび再現率、精度による評価で、それぞれ、0.298, 0.533, 0.076の値を得た。3つの提案手法を統合した手法が最も優れていることが分かった。また、この手法により得られた特許用語上位20件の中に正解が1件もない事例の数が全体の25.5%であり、この側面での評価においても、他の比較手法よりも優れていることが確認された。

参考文献

1) Fujii, A., Iwayama, M. and Kando, N.: Overview of Patent Retrieval Task at NTCIR-4, Working Notes of the 4th NTCIR Workshop, pp.225–232 (2004).
2) Fujii, A., Iwayama, M. and Kando, N.: Overview of Patent Retrieval Task at NTCIR-5, Proc. 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, pp.269–277 (2005).
3) Fujii, A., Iwayama, M. and Kando, N.: Overview of the Patent Retrieval Task at NTCIR-6 Workshop, Proc. 6th NTCIR Workshop Meeting, pp.359–365 (2007).
4) Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora, Proc. 14th International Conference on Computational Linguistics, pp.539–545 (1992).
5) Ikeda, D., Fujiki, T. and Okumura, M.: Automatically Linking News Articles to Blog Entries, Proc. AAAI Spring Symposium Series Computational Approaches to Analyzing Weblogs, pp.78–82 (2006).
6) Itoh, H., Mano, H. and Ogawa, Y.: Term Distillation for Cross-db Retrieval, Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task, pp.11–14 (2002).
7) Iwayama, M., Fujii, A., Kando, N. and Takano, A.: Overview of Patent Retrieval Task at NTCIR-3, Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task, pp.1–10 (2002).
8) 釜屋英昭，難波英樹，相沢輝昭，奥村 学：特許，論文間の引用関係を利用した論文用語の特許用語への変換，言語処理学会第12回年次大会，pp.723–726 (2006).
9) 清田陽司，黒橋聡夫，木戸冬子：自動抽出した換言表現を利用した係り受け関係の検出・自然言語処理，Vol.11, No.4, pp.127–145 (2004).
10) Mase, H., Matsubayashi, T., Ogawa, Y., Yayoi, T., Sato, Y. and Iwayama, M.: NTCIR-5 Patent Retrieval Experiments at Hitachi, Proc. NTCIR-5 Workshop Meeting, pp.318–323 (2005).
11) Mase, H. and Iwayama, H.: NTCIR-6 Patent Retrieval Experiments at Hitachi, Proc. 6th NTCIR Workshop Meeting, pp.403–406 (2007).
12) 難波英樹：論文間の引用情報を利用した関連用語の自動収集，言語処理学会第11回年次大会 (2005).
13) Nanba, H.: Query Expansion using an Automatically Constructed Thesaurus, Proc. 6th NTCIR Workshop Meeting, pp.414–419 (2007).
14) Nanba, H., Anzen, N. and Okumura, M.: Automatic Extraction of Citation Information in Japanese Patent Applications, International Journal on Digital Libraries, Vol.9, No.2, pp.151–161 (2008).
15) Nanba, H., Fujii, A., Iwayama, M. and Hashimoto, T.: The Patent Mining Task in the Seventh NTCIR Workshop, Proc. 1st International CIKM Workshop on Patent Information Retrieval (PaIR ’08), pp.25–31 (2008).
16) 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).

（担当編集委員 宮森 恒）
難波 英嗣（正会員）
1996年東京理科大学理工学部電気工学科卒業。1998年北陸先端科学技术大学院大学情報科学研究科博士前期課程修了。2001年北陸先端科学技术大学院大学情報科学研究科博士後期課程修了。同年日本学術振興会特別研究員。2002年東京工業大学精密工学研究所助手。同年広島市立大学情報科学部講師。2007年広島市立大学大学院情報科学研究科講師。現在に至る。博士（情報科学）。テキストマイニング、情報検索、自動要約、特許情報処理に関する研究に従事。言語処理学会、人工知能学会、ACL、ACM 各会員。

釜屋 英昭
2006年広島市立大学情報科学部知能情報システム工学科卒業。2008年広島市立大学大学院情報科学研究科博士前期課程修了。同年株式会社日立システムアンドサービス入社。現在に至る。修士（情報科学）。

竹澤 寿幸（正会員）
1984年早稲田大学理工学部電気工学科卒業。1989年同大学院博士後期課程修了。同年（株）日本電信通信基礎技術研究所入社。2007年広島市立大学大学院情報科学研究科教授。現在に至る。工学博士。音声対話翻訳の研究開発に従事。平成18年度電子情報通信学会 ISS 論文賞受賞。電子情報通信学会、人工知能学会、日本音響学会、言語処理学会各会員。

奥村 学（正会員）
1962年生。1984年東京工業大学工学部情報工学科卒業。1989年同大学院博士課程修了。同年東京工業大学工学部情報工学科助手。1992年北陸先端科学技术大学院大学情報科学研究科助教授。2000年東京工業大学精密工学研究所助教授。現在に至る。工学博士。自然言語処理、認知情報提示技術、語学学習支援、テキスト評価分析、テキストマイニングに関する研究に従事。人工知能学会、AAAI、言語処理学会、ACL、認知科学会、計量国語学会各会員。

新森 昭宏（正会員）
1983年京都大学理学部卒業。1990年コロラド大学大学院コンピュータサイエンス科修士課程修了。2005年東京工業大学大学院総合理工学研究科博士課程修了。1983年（株）インテック入社。現在（株）インテックシステム研究所 ICT 研究部長。博士（工学）。技術士（情報工学部門）。

谷川 英和（正会員）
1986年神戸大学工学部システム工学科卒業。同年松下電器産業（株）入社。データベースシステムの研究開発に従事。1999年弁理士試験合格。2002年よりIRD 国際特許事務所、現在に至る。博士（情報学）。弁理士、特許工学に関する研究に従事。日本知財学会等各会員。

情報処理学会論文誌　データベース　Vol. 2　No. 1　81-92 (Mar. 2009)　© 2009 Information Processing Society of Japan