優先権主張出願等のための特許書類比較ツールの提案

Patent Comparison Tool for Priority Application

IRD 国際特許事務所所長（弁理士）／株式会社アイ・アール・ディー 谷川　英和

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）に入社し、中央研究所等において、データベース管理システム等の研究開発に従事。1999 年弁理士試験合格。2002 年 1 月、IRD 国際特許事務所を開設。所長、弁理士。2003 ～ 2007 年 3 月京都大学 COE 研究員、2007 年 4 月～京都大学非常勤講師（現客員教授）、2011 年 4 月～大阪大学非常勤講師。博士（情報学）。弁理士会、日本知財学会、情報処理学会各会員。2007 年度から特許版産業日本語委員会委員。

htanigawa@ird-pat.com　TEL 06-6944-4530

IRD 国際特許事務所／株式会社アイ・アール・ディー 太田　貴久

2006 年豊橋技術科学大学大学院博士前期課程知識情報工学専攻修了。2006 ～ 2015 年同大学研究員。2016 年～ IRD 国際特許事務所。修士（工学）、言語処理学会会員、日本知財学会会員。2014 年度から特許版産業日本語委員会委員。

tota@ird-pat.com　TEL 06-6944-4530

1 はじめに

1 以上の特許出願を基礎（以下、基礎出願という。）とした国内の優先権主張出願や、外国への優先権主張出願（以下、優先権主張出願という。）において、1 以上の基礎出願の内容が、優先権主張出願に含まれているかをチェックすることは非常に重要である。もし、基礎出願の内容の一部が脱落している場合や基礎出願と優先権主張出願との整合性が取れていない場合、権利範囲が狭くなることや、権利が取得できないことがある。基礎出願のページ数が膨大な場合や、基礎出願が多数存在する場合、このようなチェックを人手で完全に行うことは容易ではない。

一方、従来、特許明細書等の作成を支援するシステム等が提案されている（例えば、文献 [1]）。また、拒絶理由通知書で示された特許明細書等の一部分と類似する単語が多い段落等を表示する装置等が提案されている（例えば、文献 [2]）。しかしながら、従来のシステム等においては、1 以上の基礎出願と優先権主張出願（国内，外国を含む）との対応を検査することができなかった。

そこで、我々は、優先権主張出願の書類に、その基礎出願の内容が含まれているかを容易にチェックできる特許書類比較ツールの開発を行った。

2 特許書類比較ツールの概要

特許書類比較ツールの概要を図 1 に示す。図 1 のように本ツールは、1 つのチェックの対象とする書類（以下、「チェック対象書類」、具体的には優先権主張出願）と、1 以上のチェック対象書類の基礎となった書類（以下、「比較対象書類」、具体的には基礎出願）を入力として受け付ける。

本ツールは、チェック対象書類と比較対象書類とを受け付けると、チェック対象書類と比較対象書類間の対応する段落を同定する。そして、その対応結果をユーザにわかりやすく表示する。ユーザへの表示方法としては、例えば、図 1 のような表形式がある。本表示方法では、対応する段落が同じ行になるように、チェック対象書類と比較対象書類の各段落を配置する。なお、本稿では対応関係を求める単位（以下、「検査単位」）を段落とするが、文などのより細かい単位や、複数段落などのより大きい単位を検査単位としても良い。

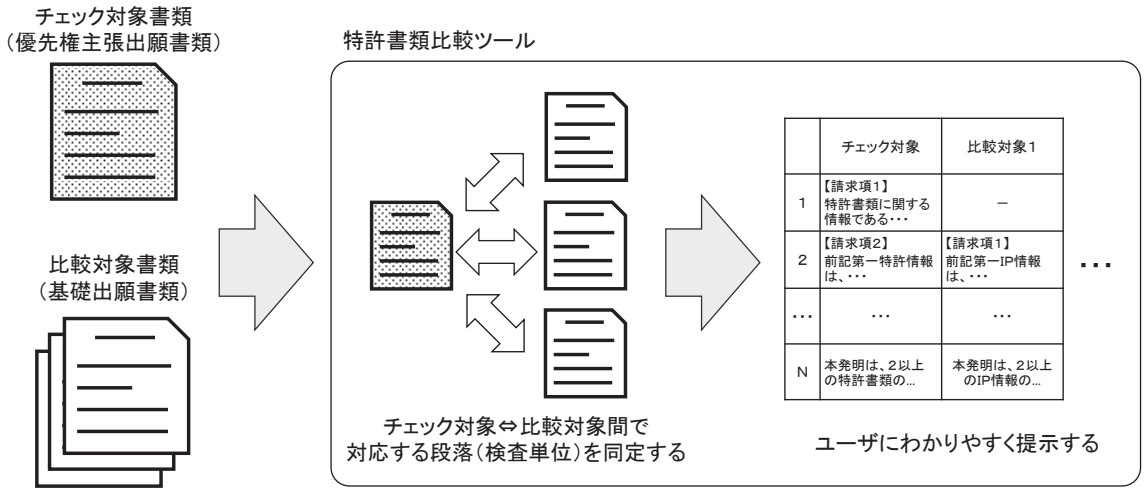


図１　特許書類比較ツールの概要

3 特許書類比較手法の概要

次に、本ツールが具体的にどのようにして、対応する段落を同定するかについて簡単に説明する。本ツールの特許書類比較手法の概要を図２に示す。

図２のように、本手法は、大きく５つのステップに分かれている。以下、各ステップについて説明する。

3.1　書類構造解析

本ツールの特許書類比較手法では、はじめに、入力された書類の構造を解析する。具体的には、日本の特許の場合、書類中に現れる隅付き括弧や、明細書中において、隅付き括弧の直後に現れる短い文字列（見出しと推定される文字列、例えば＜実施例＞等）を抽出し、その構造を解析する。本処理によって、入力されたデータを「書誌事項」、「特許請求の範囲」、「明細書」、「要約書」、「図面」等に分割し、さらに、特許請求の範囲の中の各【請求項】や、明細書の【発明を実施するための形態】といった見出しの構造を解析する。

3.2　検査単位取得

次に、構造を解析したチェック対象書類と比較対象書類のそれぞれから検査の対象とする単位を取得する。基本的には、以下のような単位を各書類から抽出する。

- 書誌事項　→　項目ごと（人は１人で１項目）
- 特許請求の範囲　→　請求項ごと
- 明細書、要約書　→　段落・図表ごと
- 図面　→　図ごと

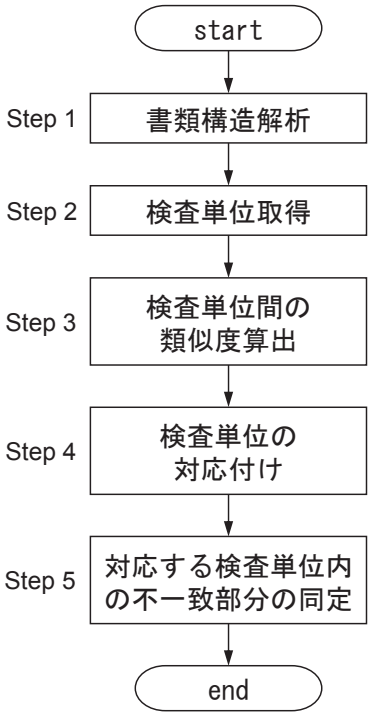


図２　手法概要

本ツールでは、上記の検査単位が基本となるが、先にも述べたように、文や複数の段落を１つの検査単位としても良い。ただし、この際、各書類の一部が複数の検査単位に重複してはならない。例えば、ある明細書に文ｓ１，ｓ２，ｓ３があるとき、１つ目の検査単位がｓ１とｓ２、２つ目の検査単位がｓ２とｓ３のように、ｓ２が重複して含まれてはならない。また、抽出した検査単位は、入力された書類上での順序を保持しているものとする。以下、簡単のために、検査単位を「段落」として説明を行う。

3.3　検査単位間の類似度算出

入力された各書類から段落を抽出後、段落間の類似度を求める。本ツールでは、チェック対象書類と比較対象書類の段落の組み合わせのすべてについて、段落間の類似度を算出する。例えば、チェック対象書類Ａと比較対象書類Ｂ１とＢ２の３つの書類を検査する場合、ＡとＢ１、ならびにＡとＢ２の２パターンの段落の組み合わせについて類似度の算出を行う。すなわち、チェック対象書類Ａが［ａ１，ａ２，ａ３］、比較対象書類Ｂ１が［ｂ１１，ｂ１２］、Ｂ２が［ｂ２１，ｂ２２，ｂ２３］という段落で構成される場合、同一書類内の段落（例えば、ａ１とａ３）や、チェック対象書類もしくは比較対象書類同士の段落（例えば、ｂ１２とｂ２３）については類似度を求めない。また、種類の異なる検査単位間での類似度の算出も行わない。例えば、異なる書誌事項間（例えば発明者と出願人）や、書誌事項とテキスト（段落）、書誌事項と図面、テキストと図面等組み合わせについては類似度の算出を行わない。

検査単位（段落）間の類似度については、検査単位の書誌事項又はテキスト又は図面がどの程度似ているかを表す数ならば、その内容は問わない。本ツールでは基本的に、書誌事項間の類似度は完全一致した場合に１、そうでない場合に０となるような類似度を、図面間の類似度は、図をヒストグラム化した後、Histogram Intersection[3] で求めた類似度を用いる。また、段落の類似度については、段落を単語の０１ベクトル（段落を「単語の異なり数」次元のベクトルとして、各次元の値はその次元に対応する単語が１回以上現れた場合は１、現れなかった場合は０とするベクトル）として、以下の式で定義されるコサイン類似度を用いる。

$$\mathrm{Sim}(a,b) = 1 - \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2} \cdot \sqrt{\sum_i b_i^2}}$$

ここで、ａおよびｂは段落を表し、ａｉおよびｂｉは、それぞれ、段落ａの単語ｉに対応する値、段落ｂの単語ｉに対応する値を表す。

3.4　検査単位の対応付け

次に、前ステップで算出した類似度をもとに、対応する段落を決定する。対応する段落を決定する最も単純な方法として、類似度が閾値以上の段落を対応すると判定する方法がある。しかしながら、特許書類では、類似した表現が繰り返し現れることがあり、このような場合、本来は対応しない段落にもかかわらず、「対応する段落」と判定される恐れがある。このような誤判定を抑制するために、本ツールでは、段落の順序関係を考慮して対応関係を求める。

段落の順序関係を考慮して対応関係を求める方法としてはグローバルアライメントを用いることも考えられる。グローバルアライメントとは、２つの配列が全体的に類似している場合に、対応関係にある配列の要素を求めるアルゴリズムである。しかしながら、図３のように、チェック対象書類と比較対象書類の複数の段落をまとめて入れ替えているような場合に、グローバルアライメントは対応できない。図３のような場合、グローバルアライメントでは［ａ１，ａ２，ａ３］と［ｂ１，ｂ２，ｂ３］の対応関係しか求めることができない。

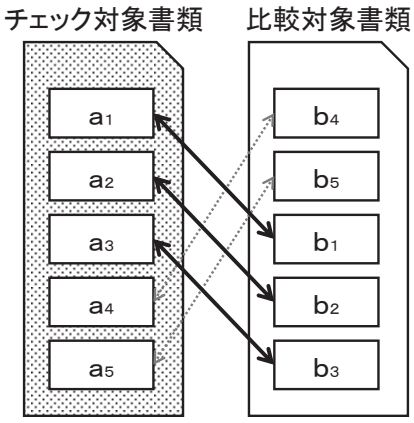


図3　グローバルアライメントでは対応できない場合

図３のような場合に対応するためには、局所的な対応関係を正しく求めることができるローカルアライメントが必要となる。ローカルアライメントを用いた場合、グローバルアライメントでは求めることができなかった図３の［ａ４，ａ５］と［ｂ４，ｂ５］の対応関係も求めることが可能となる。

本ツールでは、ローカルアライメントを求めるアルゴリズムである太田の手法 [4] を用いて段落間の対応付けを行う。太田の手法 [4] は、バイオインフォマティクス分野で用いられる Smith-Waterman アルゴリズム [5] をベースにした方法であり、学生レポートの剽窃（コピー＆ペースト）を検出するためのアルゴリズムである。簡

単に説明すると、書類Aのｉ番目の段落（ａｉ）と書類Bのｊ番目の段落（ｂｊ）が対応することを表す尺度Ｓｉ，ｊを以下のように定義する。

$$S_{i,j} = \max\begin{cases} 0, \\ \mathrm{Sim}(a_i, b_j) - \alpha + S_{i-1,j}, \\ \mathrm{Sim}(a_i, b_j) - \alpha + S_{i,j-1}, \\ \mathrm{Sim}(a_i, b_j) - \alpha + S_{i-1,j-1} \end{cases}$$

ここで、αは段落間の類似度の閾値を表す。

この値が最も大きくなるｉ，ｊの組み合わせを求め、その値が0より大きい場合に、ａｉとｂｊが対応関係の末尾であると判定する。アルゴリズムの詳細は文献 [4] を参照されたし。なお、今回、閾値αについては、前ステップで求めた段落間の全類似度に対して判別分析法（大津の２値化）を適用して求めた。

3.5　対応する検査単位内の不一致部分の同定

段落間の対応関係は、前ステップまでの処理で求めることが可能だが、本ツールでは、さらに対応すると判定された段落について、具体的に段落内のどの部分が異なるかを検出する。具体的な不一致部分を検出し、それを提示することで、前ステップで対応関係の導出を誤った場合でも、ユーザはその誤りに容易に気付くことができる。

段落（テキスト）の不一致部分の検出については、先に説明したグローバルアライメント手法の１つであるレーベンシュタインの編集距離を求める手法で対応付けを行い、不一致部分を検出する。

一方、図面については、サイズを合わせたのち、その差分をとり、差分が閾値を超えた領域に色を付けて出力する。

4　出力のイメージ

最後に、上記の特許書類比較方法を用いて２つの特許書類を比較した際の出力イメージを図４に示す。

図４のように、本ツールは、チェック対象書類と比較対象書類を対応関係にある段落が同じ行に配置されるように出力する。さらに、対応関係が見つからなかった段落や、段落内のチェック対象書類と異なる部分に色を付けて出力するため、ユーザは２つの特許書類の対応関係を容易に把握することができるため、記載漏れ等を防止することが可能となる。

5　おわりに

本稿では、１以上の基礎出願に基づいて優先権主張出願を作成するとき、その基礎出願が優先権主張出願に含まれているかを容易にチェックするための特許書類比較ツールの概要について紹介した。基本的な機能としては、本稿で紹介した内容で十分であるが、本ツールにはいくつかの課題がある。

本ツールの適用対象の典型的な例として、ある基礎出願を別の言語に翻訳して外国へ優先権主張出願する場合がある。本稿で紹介した特許書類比較方法は同一言語で記述された書類を前提としているため、そのままでは「全段落が不一致」と判定されてしまう。そこで、チェック対象書類と比較対象書類のいずれかを機械翻訳し、翻訳結果を用いて対応関係や不一致部分を求める機能が必要である。

また、現在のツールの図面に対する処理は非常にシンプルな方法を用いて、類似度の算出や不一致部分の同定を行っている。より有用なツールとするために、今後、ＯＣＲを用いて図面内の文字列を解析し、図面内の文字列も考慮した類似度や不一致部分の表示を行う必要がある。

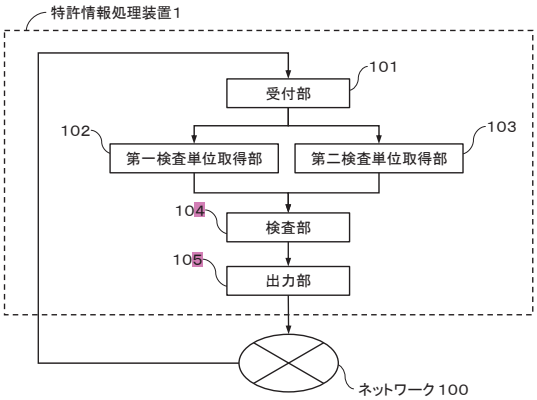
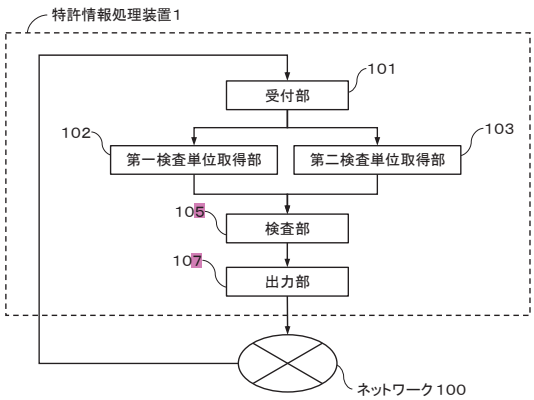
チェック対象書類	比較対象書類	
	段落	判定
【請求項1】 特許書類に関する情報である第一特許情報と、当該第一特許情報に基づいて作成された特許書類に関する情報である第二特許情報を受け付ける受付部と、 前記第一特許情報の一部である第一検査単位を各第一特許情報から取得する第一検査単位取得部と、 前記第二特許情報の一部である第二検査単位を各第二特許情報から取得する第二検査単位取得部と、 前記第一検査単位と前記第二検査単位との対応を検査し、検査結果を取得する検査部と、 前記検査部が取得した検査結果を出力する出力部とを具備する特許情報処理装置。	【請求項1】(類似度0. 98) 特許書類に関する情報である第一特許情報と、当該第一特許情報に基づいて作成された特許書類に関する情報である第二特許情報を受け付ける受付部と、 前記第一特許情報の一部である第一検査単位を各第一特許情報から取得する第一検査単位取得部と、 前記第二特許情報の一部である第二検査単位を各第二特許情報から取得する第二検査単位取得部と、 前記第一検査単位と前記第二検査単位との対応を検査し、検査結果を取得する検査部と 前記検査部が取得した検査結果を出力する出力部とを具備する特許情報処理装置。	A
【請求項2】 前記第二特許情報は、 前記第一特許情報に対応する特許の優先権主張出願の特許書類に関する情報である請求項1記載の特許情報処理装置。	【請求項3】(類似度1. 00) 前記第二特許情報は、 前記第一特許情報に対応する特許の優先権主張出願の特許書類に関する情報である請求項1記載の特許情報処理装置。	A
	【請求項2】(類似度0. 95) 前記第一特許情報は、 前記第二特許情報に対応する特許の優先権主張出願の特許書類に関する情報である請求項1記載の特許情報処理装置。	A
【請求項3】 前記第一特許情報は、第一言語で記述された情報であり、 前記第二特許情報は、第二言語で記述された情報であり、 前記第二特許情報は、 第一特許情報を翻訳した特許書類に関する情報である請求項1記載の特許情報処理装置。	(対応なし)	N
	【請求項4】(対応なし) 前記第一特許情報を第二言語に機械翻訳し、第一翻訳情報を取得する機械翻訳部をさらに具備し、 前記第一検査単位取得部は、 前記第一翻訳情報から第一検査単位を取得する請求項1記載の特許情報処理装置。	F
⋮	⋮	⋮
【背景技術】【0002】 優先権主張出願や、ある出願を別の言語に翻訳して外国へ出願するとき等の、ある特許書類を基礎として別の特許書類を作成するとき、その基礎となった書類の内容が、作成した特許書類に含まれているかをチェックすることは非常に重要である。もし、基礎となった書類の内容の一部が含まれなかった、もしくは新規に内容を追加してしまった場合、権利範囲が狭くなることや、権利が取得できないことがある。基礎となる書類のページ数が膨大な場合や、基礎となる書類が多数存在する場合、このような検査を人手で完全に行うことは困難である。	【背景技術】【0002】(類似度0. 90) 優先権主張出願等の、ある特許書類を基礎として別の特許書類を作成するとき、その基礎となった書類の内容が、作成した特許書類に含まれているかをチェックすることは非常に重要である。もし、基礎となった書類の内容の一部が含まれなかった、もしくは新規に内容を追加してしまった場合、権利範囲が狭くなることや、権利が取得できないことがある。基礎となる書類のページ数が膨大な場合や、基礎となる書類が多数存在する場合、このようなチェックを人手で完全に行うことは困難である。	B
⋮	⋮	⋮
【図1】  	【図1】(類似度0. 85)  	B
⋮	⋮	⋮

図4　特許書類比較ツールの出力イメージ

参考文献

[1] 谷川英和，田中克己，”３種類の特許部品データベースに基づく特許明細書生成エンジンの構築”，情報処理学会論文誌：データベース，Vol.47，No. SIG 8(TOD30)，pp.90-102, 2005.

[2] アイビーリサーチ有限会社，知的財産管理装置，特開 2012-242879 号公報，2012.

[3] Swain, M., and D. Ballard, “Color indexing,” International Journal of Computer Vision, 7, 1 pp. 11-32, 1991.

[4] 太田貴久，増山繁，" 学生レポート採点支援のためのレポート類似部分発見手法 "，信学技報，NLC2005-112, pp. 37-42，2006.

[5] Smith, Temple F.; and Waterman, Michael S. "Identification of Common Molecular Subsequences", Journal of Molecular Biology 147, pp.195-197，1981.