文書内の言語構造を利用した
特許文書分類・検索技術の研究

間瀬　久雄

目 次

図表一覧

第１章　序論

1.1　研究の背景

1.1.1　社会的背景

近年の企業競争の急速なグローバル化に伴い，多くの企業では熾烈な競争に打ち勝つべく，特許をはじめとする知的財産権の戦略的な取得・管理・活用に力を入れるようになってきた．すなわち，**図 1.1** に示すように，単に製品の販売やサービスの提供だけでなく，製品やサービスの核となる高付加価値技術を知的財産として行使する権利を保有することにより，他社との競争を優位にしようという動きが一段と鮮明になってきている．

このような知的財産権に対する企業の積極的な動向を背景として，2002 年 7 月に日本政府は「知的財産戦略大綱」[1]を発表している．本大綱では，「日本の産業競争力低下への懸念と，知的創造サイクルの確立の必要性」を現状課題として踏まえ，「知的財産をもとに，製品やサービスの高付加価値化を進め，経済・社会の活性化を図る『知的財産立国』を実現する」と明記されている．そして，その戦略として，知的財産の「創造戦略」，「保護戦略」，「活用戦略」，「人的基盤の充実」の四つが掲げられている．このうちの「保護戦略」の具体策の一つとして，「迅速かつ的確な知的財産権の審査・審判」が挙げられており，審査体制の整備や国際的協調を含む総合的な対策が必要であると示されている．

知的財産権の大部分を占める特許は毎年 42 万件前後出願されており，1990 年における特許の電子出願の施行以降だけでも，累計で 600 万件以上の特許文書が電子文書の形で蓄積されている．また，2001 年に従来の特許審査請求期間（出願日を起点とした，特許の権利化に係る審査の実施を請求できる期間）が 7 年から 3 年に短縮されたことに伴い，2004 年における特許審査請求件数は，前年比 32%増の約 33 万件と急増している．その結果，審査請求から審査開始までの審査待ち期間が 26 ヶ月にまで長くなっており，特許庁における特許審査業務を圧迫している[2]．特許は網羅する技術分野が非常に広く，審査には技術分野の専門知識及び特許文書検索ノウハウを要するため，審査官の増員も容易ではない．

以上の社会的背景から，特許審査期間の抜本的短縮及び審査精度の一層の向上が，企業や諸外国から強く要求されてきている．そしてその一環として，特許庁における審査業務形態を踏まえた特許審査の支援や，審査ノウハウの蓄積・共有など，特許審査業務の効率

---

[1] 知的財産戦略大綱 http://www.kantei.go.jp/jp/singi/titeki/kettei/020703taikou.html

[2] 特許行政年次報告書 2005 年版 http://www.jpo.go.jp/shiryou/toushin/nenji/nenpou2005_index.htm

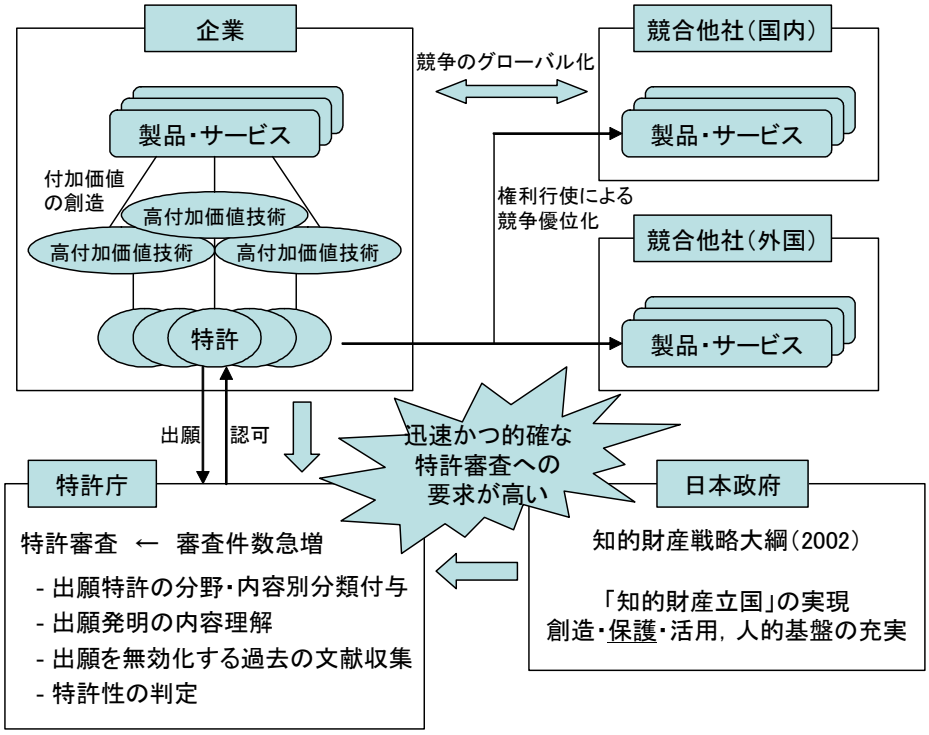


図 1.1　企業における知的財産権（特許）の取得と活用

向上に寄与する計算機支援システムの実現への要求が一層高まってきている．具体的には，出願特許の内容を理解したり，発明内容毎に分類整理したり，記載された発明内容を無効化する過去の文献を収集したりする知的作業を，多角的かつ柔軟に支援する特許審査支援システムである．

一方，知的財産権の戦略的活用を模索している企業内においても，製品やサービスに係る特許の戦略的な取得・活用と，特許管理の低コスト化が大きな課題となっており，特許の戦略的活用の促進を支援する計算機システムへの期待が高まっている．具体的には，出願された特許をその企業独自の戦略軸に沿って分類整理したり，ある製品・サービスに対する自他社技術を分析し，その優劣を評価することにより，今後の企業戦略を立案したりする知的作業を支援する特許分析・管理システムである．

特許庁での特許審査業務及び企業での特許戦略立案業務に共通する作業で，業務の遂行に最も影響を与える作業は，大量の特許文書の中から，ある特定の条件を満たす特許文書を収集する特許文書検索作業である．特に特許庁での特許審査業務では，この検索作業の良し悪しによって企業間の訴訟裁判に発展する恐れがあるため，検索作業には緻密な検索戦略と結果の正確さが常に要求される．

文書検索の高精度化に対する要求は，特許文書を対象とした検索だけに限らない．新聞記事や科学技術文献，設計文書，Web ページ，電子メールなどの文書の中から所望の文書だけを効率良く取得したいという要求は，1990 年代後半からのインターネット及び PC の普及による文書情報量の爆発的増大に伴い，日増しに高くなっている．特に Web ページ検索においては，不特定多数の人々によって作成・発信される 110 億ページもの文書の中から，ユーザの多種多様な目的の遂行において必要な文書を，迅速かつ的確に検索する技術が要求される．ある条件を満たす文書を「高速に」検索する技術は，著名な Web 検索エンジンに見られるように実現されてきているが，「的確に」検索する技術については発展途上にあり，ユーザの満足するレベルに達していないのが現状である．

1.1.2 技術的背景

現在，数多くのユーザに利用されている文書検索システムにおいて最も主流な検索方式は，キーワード（以下，本研究では「ターム」と呼ぶ）による論理式検索である．すなわち，自分の欲しい情報を AND/OR/NOT などの論理演算子を用いたタームの論理式で表現し，その論理式が真となる文書を検索する方式である．本方式は，ターム間の論理関係を明確に定義できる，論理式に合致する文書を漏れなく検索できる，などの利点を持つ反面，適切なタームを選択して論理式を組み立てるのにノウハウを必要とする，検索結果の順位付けが困難である，などの欠点がある．

これらの欠点を解消すべく，所望の情報に関連するタームを提示したり，検索結果集合におけるタームまたは文書間の意味的関連をビジュアル表示したりする支援機能を持つ検索システムが出現してきている．また，ユーザの欲しい情報を文章で入力し，その文章に内容が近い文書を検索する類似文書検索（自然言語文検索）システムや，質問文に対する回答を出力する質問応答（Q&A）システムも徐々に普及し始めてきている．特に類似文書検索は，情報検索に対する要求を文章で入力できるため，論理式検索に比べ，検索条件を作成する手間を大幅に削減できるという利点がある．また，論理式検索で使われる論理式に比べて，自然言語文からは出現頻度や構文など，文書検索で有用な手掛かり情報をより多く抽出できるため，検索結果の順位付けによる検索の高精度化が可能である．

しかし，これら種々の検索システムが普及してきた現在でも，人々は依然として情報を探すのに多大な時間を費やしている．その理由の一つとして，上記のどの検索方式を適用してみても検索精度がユーザにとってまだ十分に高くないために，ユーザは検索を試行錯誤的に何度も繰り返さざるを得ないでいることが挙げられる．したがって，ユーザが入力する限られた検索条件を手掛かりとして所望の情報を高精度に検索することは，ユーザの検索作業全体の効率向上には欠かせない最も重要な技術課題である．

検索精度を向上させることで検索の試行錯誤を極力少なくし，検索作業の効率化を実現するために解決すべき技術課題は多い．すなわち，ユーザの欲しい情報をどのように表現するか，検索対象となる文書集合に記載される内容をどのように定式化するか，両者をどのように照合して所望の情報か否かを判定するか，文書固有の手掛かりや分野固有の特性を検索にどのように取り込むか，ユーザからのフィードバック情報を以降の検索にどのように反映させるか，などである．これらの課題を克服し，ユーザの作業負担を軽減する文書検索システムを実現することで，初めて検索作業の効率を飛躍的に高めることができる．

1.2　研究の目的と位置付け

本研究では，世の中に流通している多種多様な文書情報の中から特許文書を研究対象として採り上げる．大量の特許文書の中から所望の特許文書を迅速かつ的確に検索する方式を実現することにより，特許庁での審査業務及び企業での特許戦略立案業務の効率改善と質の向上を目指す．本節では，特許庁での特許審査業務に焦点を絞り，本研究の目的及び位置付けについて，現状の特許審査業務に対応付ける形で具体的に述べる．

特許審査業務は**図 1.2** に示すように大きく以下の三つの作業プロセスからなる[3]．

(1) 方式審査

出願された特許文書に記載漏れはないか，発明内容が公序良俗に反する内容でないかなどをチェックする．

(2) 分類付与

出願特許文書に記載された発明内容に応じて，予め定義された特許分類体系の中から適切な分類を割り当てる．特許文書検索時にこれらの分類情報を検索条件として活用することにより，検索対象文書集合を絞り込むことが可能となる．

(3) 特許性審査

出願特許文書に記載された発明内容が特許として成立するか否かを判定する．すなわち，過去に同じ発明内容が存在していないかを検索によってチェックする．

上記作業プロセスのうち，(1)方式審査及び(2)分類付与は，出願特許文書が公開特許公報として一般に広く公開される前に行われる作業である．現在，出願から公開までに要する作業期間は 1 年強であるが，この大半は分類付与作業に費やされている．特許分類体

[3] 特許庁ホームページ　http://www.jpo.go.jp/tetuzuki/index.htm

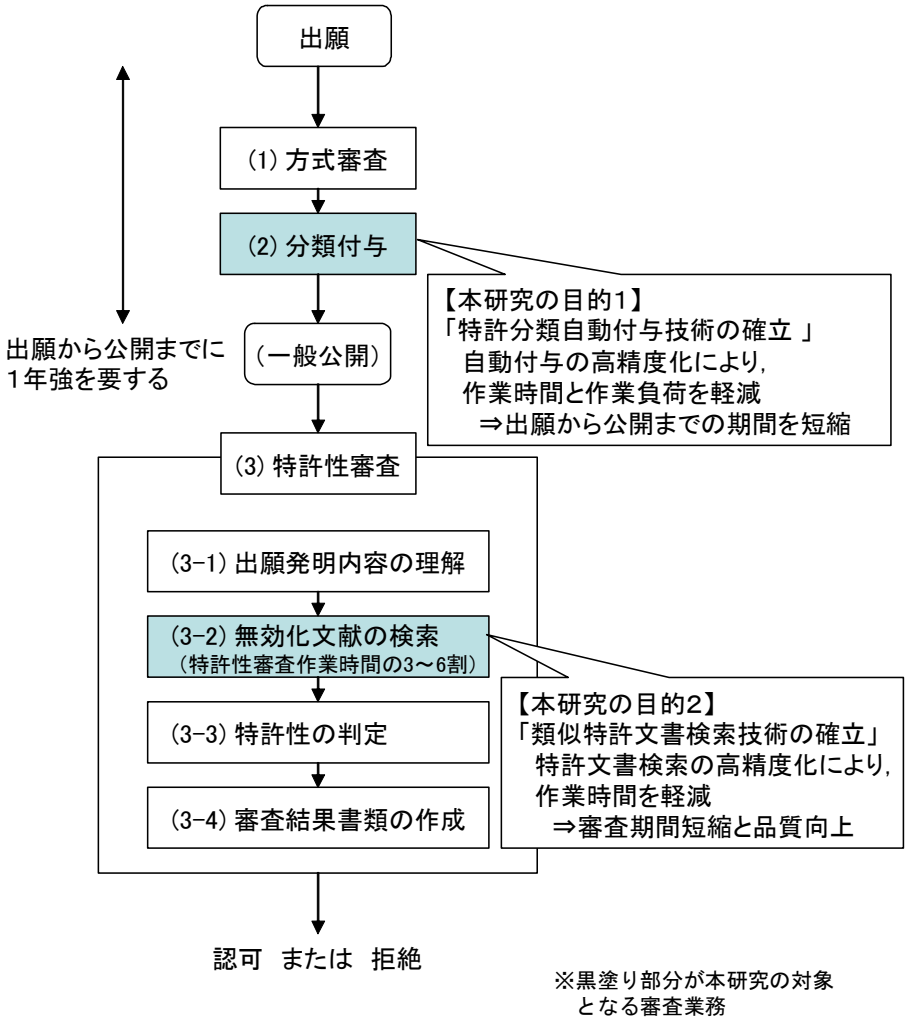


図 1.2　特許審査業務の流れと本研究の目的及び位置付け

系は他に類を見ないほど大規模である．すなわち，世界標準である国際特許分類（IPC; International Patent Classification，約 7 万分類），IPC の詳細分類で日本独自の分類体系である FI（File Index，約 19 万分類），IPC とは独立に定義された日本独自の観点別分類体系である F ターム（約 30 万分類）がある[4]．現状ではこれらの分類を数百人の分類付与専門家が人手で付与している．そのため，多大な人件費と作業時間がかかる，分類付与が担当者のスキルに依存するため付与結果が均一でない，分類体系が改正された時に過去の特許に改正後の分類を付与し直すコストが膨大である，などといった問題が生じている．

[4] 特許庁パテントマップガイダンス　http://www5.ipdl.ncipi.go.jp/pmgs1/pmgs1/pmgs

一方，(3)特許性審査は，更に以下の 4 段階の作業プロセスからなる．

(3-1) 出願発明内容の理解

審査対象となる特許文書を熟読し，記載された発明内容について十分に理解する．

(3-2) 無効化文献の検索

特許文書に記載された発明内容と同一あるいは類似する発明内容について書かれた過去の特許文書や論文などの文献を収集する．

(3-3) 特許性の判定

収集された文献の内容と審査対象の発明内容を比較し，特許として認めるかを判定する．

(3-4) 審査結果書類の作成

特許として認めるか否かの判定結果を公式文書としてまとめる．

このうち，(3-2)無効化文献の検索に要する時間が特許性審査作業全体の 3 割から 6 割を占めると言われており，この作業の抜本的な効率化が要求されている．具体的には，検索精度の向上のほか，検索条件式の作成支援や検索結果の妥当性判定支援の実現である．

特許審査業務に係る以上の現状を鑑み，本研究では以下の二つの特許文書処理技術を確立することにより，特許審査期間の短縮及び審査品質の向上，作業者の労力軽減を実現することを目的とする（特許審査業務における本研究の位置付けを図 1.2 に併せて示す）．

【本研究の目的 1】特許分類自動付与技術の確立

特許文書に記載された発明内容を自然言語解析して発明の特徴を特定し，既存の特許分類体系の中から適切な分類を選択付与する技術である．この技術によって分類付与に係る作業時間及び作業負荷を改善し，出願から公開までに要する期間を短縮する．

【本研究の目的 2】類似特許文書検索技術の確立

技術分野が多岐に渡る数百万件規模の特許文書集合の中から，入力指定された特許文書に記載された発明内容を無効化する過去の特許（以下，「無効化特許」と呼ぶ）をノイズや漏れなく高精度に検索する技術である．この技術によって検索作業時間を改善し，審査期間を短縮するとともに，審査の質を向上させる．

1.3　特許文書の言語構造と特許分類体系

ここでは，本論文で述べる内容の理解を助けるために，特許文書の構成，特許請求項の文章構造と記載方法の特徴，特許分類体系の構成及び特徴について解説する．

1.3.1　特許文書の構成

出願された特許は特許庁で審査される．特許出願後，出願日や分類などの属性データが専門家によって付加され，公開特許公報として一般に公開される．

出願した発明が特許として認められるかの判断基準として，大きく「新規性」と「進歩性」の 2 点がある．「新規性」は，これまでに公にされていない新しい発明であるか否かという基準で審査される．「進歩性」は，その技術分野の人が従来技術から容易に考え付くことができない発明であるか否かという基準で審査される．したがって，特許文書を執筆する際には，上記 2 点を意識しなければならない．

特許文書は半構造化された文書である．日本の特許文書は容易に記載箇所を特定できるように以下のタグを備えている（2005 年現在）．タグ名称は括弧【】で囲んで記載するのが原則である．

【発明の名称】

発明の内容を端的に記載したもので文書タイトルに相当する．実際には，「文書検索装置」「電子レンジ」など，発明の及ぶ対象物を記載することが多い．

【請求項】

一つの請求項の中で一つの発明内容を一文で記載する．請求項は番号付けられて記載される．請求項は階層的になっていることが多く，他の請求項の発明内容に従属しない独立請求項と，他の請求項の発明内容に依存してその詳細について記載している従属請求項に分けられる．

【技術分野】

発明が適用される技術分野・対象物に関して記載する．

【従来技術】

発明に関連した，既に開示されている技術の内容に関して記載する．

【発明が解決する課題】

上記従来技術が抱える課題について言及する．

【課題を解決する手段】

発明が上記の課題をどのように解決するか，その手段について記載する．請求項の記載内容と重複することが多い．

【発明の実施の形態】

発明が実際に実現できることを，図面を用いて具体的かつ詳細に示す．複数の実施形態が記載されることも多い．

【発明の効果】

発明がもたらす効果について簡潔に記載する．

【図面の簡単な説明】

説明に用いられる各図面の記載内容について簡潔に記載するもので，図面タイトルに相当する．

【符号の説明】

図面の構成要素に付与される符号が何を示しているかを端的に記載する．

【要約】

発明の目的・課題・手段について，400 字程度で記載する．

特許文書に関するその他の特徴としては，以下の三つが挙げられる．

(1) 不特定多数の著者によって執筆されている

非常に多くの発明者が特許文書を執筆しているので，その文章記述スタイルや文章長，使用語彙（異表記・同義語含む）はさまざまであり，これが第 2 章以降で述べる特許分類自動付与や類似特許文書検索の精度を低下させる大きな要因の一つとなっている．

(2) 技術分野が広い

特許は，化学，医学，コンピュータ，食物，ビジネスモデルなど，広い技術分野を網羅しており，執筆スタイルも分野毎で異なっている．例えば，化学分野の特許文書では発明の新規性を述べるのに化学式を頻繁に使用し，機械分野の特許文書では図面を使用する．この特徴は，第 2 章以降で述べるターム抽出や重み付けのアルゴリズムに大きく影響する．

(3) 発明者と出願人と執筆者の関係

発明者はその発明を考案するのに貢献した人々を指す．また，出願人はその発明を特許として権利化することを申請する人または組織である．大部分の特許文書は発明の考案に最も寄与した筆頭発明者が執筆するが，特許事務所や企業内の知財部の特許専門家によって執筆されることもある．

1.3.2　特許請求項の文章構造

特許文書を構成するタグの中で，発明内容を最も端的に表すのは【請求項】タグである．ジェプソン型と呼ばれる典型的な請求項構造は，図 1.3 に示すように以下の構成要素からなる．

(1) 前提部分

この構成要素には，発明に関する前提条件や既知の技術（従来技術）が記載される．前提部分のほとんどは各請求項の前半部分に位置し，「～において」「～であって」などの付属語表現を伴って記載される．前提部分は必ずしも請求項に存在している訳ではない．予備調査結果によれば，【請求項 1】文章の約 52%にのみ，前提部分が含まれている．

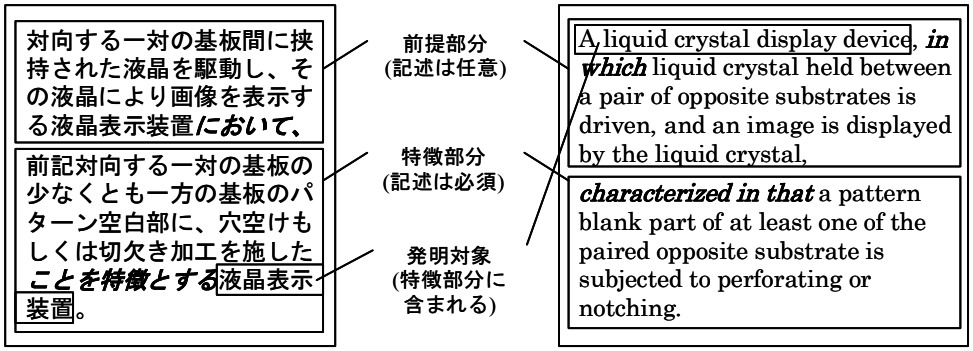
(2) 特徴部分

この構成要素には，発明の新規性または進歩性が記載される．この構成要素では，「～を特徴とする」「～を備えた」「～からなる」といった特許請求項固有の表現を伴い，これに続いて発明対象が記述される．特徴部分は，発明の構成要素（手段）を列挙したものや，発明の処理内容（ステップ）を記載したもの，ある特定の構成要素または処理内容に関する特徴を記載したものなどがある．

請求項の記載に関する上記以外の特徴としては以下が挙げられる．

(1) 記載内容の抽象化

請求項や発明の名称では，その発明を競合他社に容易に発見されないように，その内容を意図的にぼかして抽象的に記述することがある（例：「情報処理装置」）．



(NTCIR-4特許検索タスクフォーマルラン課題#23より引用)

図 1.3　請求項の文章構造

(2) 指示語の不使用

請求項では，「これ」「それ」「この」などの指示語は，意味の曖昧性を避けるために使用できない．代わりに「前記～」「当該～」といった言葉を使って，前出した単語を繰り返して使用することが多い．

1.3.3 特許分類体系

出願特許に付与される特許分類体系は，大きくは**図 1.4** に示す 2 種類に分かれる．

(1) IPC，FI

国際特許分類（IPC；International Patent Classification）は世界共通の特許分類体系である．IPC は階層構造をなしている．最上位分類であるセクション（8 分類），メイングループ（約 7,300 分類），サブグループ（約 61,500 分類）の 5 階層からなる木構造をなしている．IPC は体系の見直し・改訂が随時なされており，少しずつ体系が変化している．IPC を活用することにより，外国出願された特許を横断的に検索することが可能となる．FI（File Index）は IPC を更に詳細化した日本独自の分類体系であり，約 19 万分類からなる．IPC では区別できないレベルの発明内容でも，FI を活用することにより検索結果を絞り込むことが可能となる．

(2) F ターム，テーマ，審査室

IPC が出願特許を整理するために付与される分類であるのに対して，F タームは検索用インデクスとして使われる日本独自の分類である．F タームは，発明内容をさまざまな観点から分類するものである．着目する観点は，目的，用途，適用対象など，技術分野に依存しない観点もあれば，技術分野に特化した観点もあり，多種多様である．F タームの総

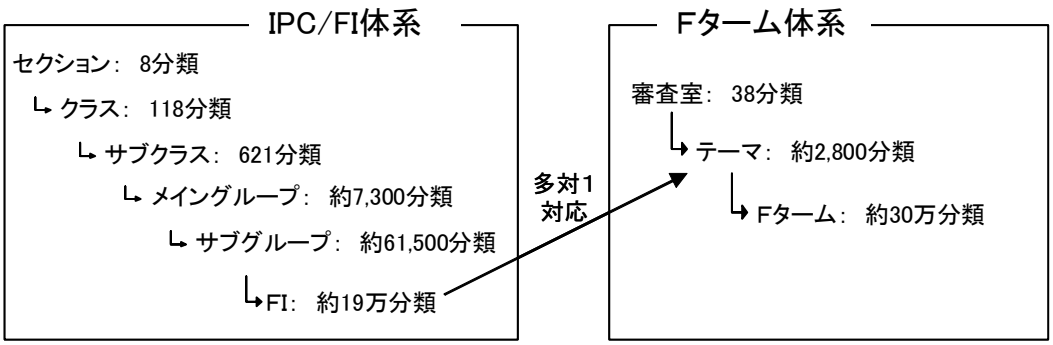


図 1.4 日本の特許分類体系

表 1.1　特許分類体系（テーマ）のサンプル（建築物関連のテーマ）

テーマ	テーマ名称	テーマ	テーマ名称
2E001	建築環境	2E121	建築構造一般
2E002	耐力壁、カーテンウオール	2E125	建築構造の接合一般
2E003	足場	2E130	組積造壁;現場打壁
2E011	開口に固定される戸、窓等の枠	2E131	薄板耐力壁;間仕切り壁
2E013	出窓，玄関ユニット及び建付調整	2E134	屋根構造
2E014	ウィング枠及びウィングの配置	2E135	その他の屋根ふき
2E015	伸縮扉、および回転扉	2E137	乗り物を格納するための建築物
2E016	ガラス板等の固定及び戸板	2E139	異常な外部の影響に耐えるための建築物
2E020	面格子、雨戸枠、戸袋	2E140	塔;農工業用築物;大型貯蓄容器の建設
2E024	建築物の仕上げ用具	2E141	テント・膜構造
2E025	居住または事務用建築物	2E142	囲い
2E026	公共建築物	2E150	建築現場における取りはずす型枠、補助部材
2E030	蝶番	2E161	建築用ブロック部材
2E032	蝶番の付属品;滑動ウイング用の付属品	2E162	建築構造用パネル
2E034	ウイング用の支持装置	2E163	建築用棒状部材
2E035	枠またはウィング用のかど部の接続	2E164	建築物の補強部材
2E036	戸・窓の密封・換気・特殊装置	2E172	コンクリート打設にともなう現場作業
2E037	網戸	2E173	現場におけるコンクリートの補強物挿入作業
2E038	門	2E174	現場における建築要素の搬送及び組立作業
2E039	特殊ウィング	2E175	現場における安全保護手段
2E042	シャッタ等の閉鎖部材	2E176	既存建築物への作業
2E043	ブラインド	2E178	戸または窓の固定装置
2E044	はしご	2E179	金庫
2E050	ウイング開閉機構;ウイング用付属品	2E181	野外携帯装備品
2E052	ウイング用動力操作機構	2E182	カーテン・垂れ幕等戸・窓の付帯設備
2E101	建築物の階段	2E184	人命救助
2E103	床構造	2E185	呼吸装置;防護
2E104	天井構造	2E189	防災
2E105	建築物の日除け・日覆い	2E191	消火剤;有害な化学剤の無害化
2E108	屋根ふき・それに関連する装置または器具	2E192	構造要素一般
2E110	壁の仕上げ	2E220	床の仕上げ
2E111	垂直ダクト;みぞ;建築仕上のその他の部分	2E250	錠;そのための付属具

数は約 30 万分類であり，出願特許 1 件当たりに付与される F ターム数は IPC や FI に比べてはるかに多く，平均すると 20 個前後付与される．テーマは F タームの上位分類に相当するものであり，約 2,800 分類からなる．特許庁の審査官や分類付与担当者はテーマ毎にアサインされている．審査室はテーマの上位の分類であり，約 40 分類からなる．テーマがどのくらいの粒度で定義されているかを，**表 1.1** にサンプルとして示す．

IPC/FI 体系とテーマ/F ターム体系は対応付けられている．すなわち，FI とテーマの間には多対 1 の関係にあり，FI が決まれば対応するテーマが一意に決まる．また，二つの分類体系は技術分野によって使われ方が異なっている．ある技術分野における特許検索ではテーマ/F タームが頻繁に使われるが，別の技術分野では IPC/FI が多用される（F タームが定義されていない技術分野も存在する）というように分野依存性が強い．

1.4　研究の基本課題と解決アプローチ

本研究では特許文書をテキストとして捉え，自然言語処理及び統計処理に基づくテキスト解析によって抽出されるタームを用いて特許文書に分類を付与したり，類似する特許文書を検索したりするという研究アプローチを前提とする．分類付与技術と類似文書検索技術には多くの技術的共通点が存在するため，両技術に係る基本課題も共通点が多い．

1.4.1　特許分類付与研究の基本課題と解決アプローチ

本研究で想定する特許分類付与作業の流れを**図 1.5** に示す．特許文書のように分類の付与精度が特許文書検索精度に大きく影響を与える場合，分類付与の高精度化は必須の要件となる．しかし，特許分類体系は大規模かつ多岐に渡っているため，人手を介さずに計算機システムが全自動で必要十分な分類を付与することは不可能に近い．

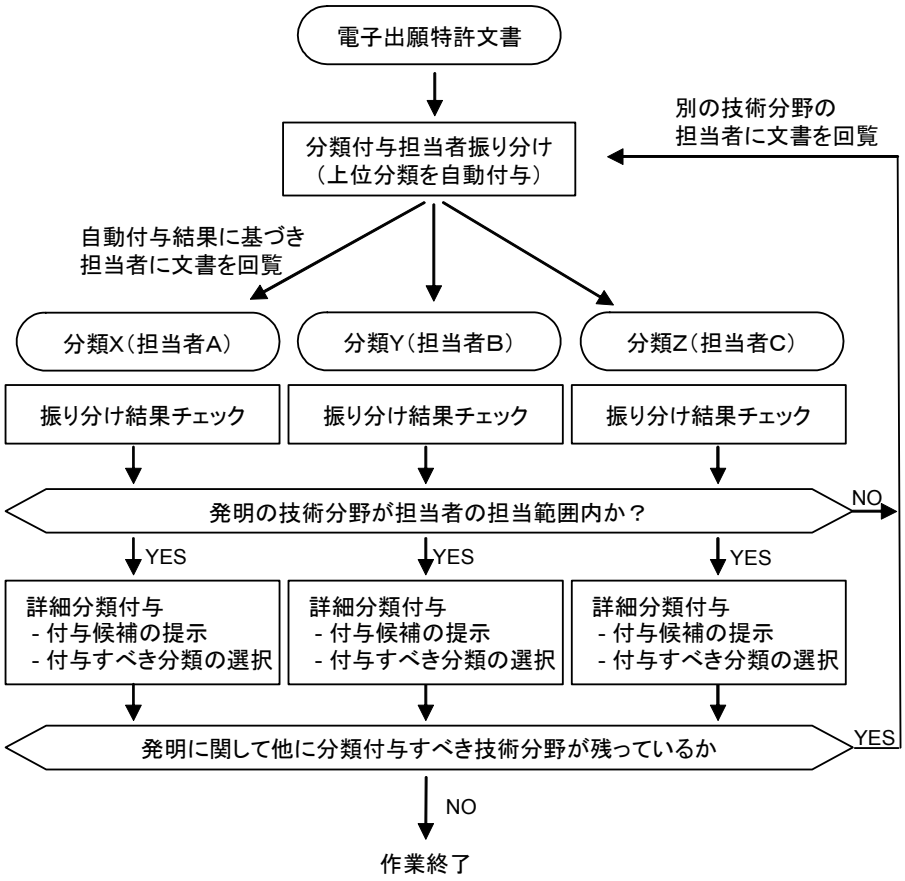


図 1.5　想定する特許分類付与作業の流れ

そのため，計算機システムによる自動付与結果を人手でチェックし，誤った分類が付与されている場合はそれを修正するという作業プロセスを組み込まざるを得ない．この際，誰がどの特許文書の付与結果をチェックするのかということが運用上の問題となる．

特許庁における分類付与作業では，特許分類体系の一つであるＦタームの上位分類にあたるテーマ（約 2,800 分類，図 1.4 参照）に対応して分類付与担当者が割り当てられている．したがって，この上位分類（テーマ）の付与が自動化できれば，その特許文書を適切な詳細分類付与担当者に自動的に振り分けることができる．また，その担当者が詳細分類（FI，F ターム）を付与する際に，付与すべき詳細分類の候補を計算機システムが分類自動付与結果として担当者に提示することにより，担当者はより効率良く分類を特定できる．

そこで以下では，分類付与担当者を割り当てるための上位分類の自動付与と，その担当者が詳細分類を付与する作業を支援する詳細分類付与支援という二つの側面から，分類付与研究の基本課題及び解決アプローチについて考察する．

分類付与研究の基本課題及び解決アプローチを**表 1.2** にまとめる．これらのうち，本研究では以下の四つの基本課題を扱うこととする．

［課題 1］特許文書に記載される発明内容をどのように記述するか？

特許文書に分類を自動付与するためには，その特許文書に記載される発明内容を解析して計算機に理解できる形式で記述する必要がある．自然言語処理分野では，入力された文章の意味を把握するために，その文章を解析して中間表現に変換するアプローチがあり，機械翻訳や対話システムなどで採用されている．一方，文書検索分野では，入力された文章の意味を，文章中の重み付きタームの集合として記述するアプローチが主流である．

［課題 2］特許文書に記載される発明内容をどのように特定するか？
（発明内容を表すタームをどのように抽出するか）

一般に，文書内容を特徴付ける重要なタームを機械的に抽出する際に用いられる観点としては，タームの出現頻度（繰り返して使用されているか），一般性（どのくらい多くの文書で使われているか），品詞，構文的位置付け（主語，目的語，述語など），主題性（そのタームに関して述べた文章であるか），出現位置（タイトル，要約，本文）などがある．

［課題 3］各分類が網羅する技術分野範囲をどのように定式化するか？

特許文書に記載された発明内容がどの分類に属するかを判定するためには，各分類が網羅する技術分野がどこからどこまでかを計算機に把握させる必要がある．ここで，どのようなリソースから各分類の網羅範囲を特定するのかが課題となる．

表 1.2　特許分類付与研究の基本課題と解決アプローチ

#	基本課題	本研究での解決アプローチ
1	特許文書に記載される発明内容をどのように記述するか？	・重み付きターム集合として発明内容を記述する
2	特許文書に記載される発明内容をどのように特定するか？	・特許文書の構成及び記載文章の構造に着目して発明内容を端的に表すタームを抽出する ・発明の構成要素よりもその発明が適用される対象物または技術分野に係る記載箇所を重要視してターム重みを算出する（上位分類付与）
3	各分類が網羅する技術分野範囲をどのように定式化するか？	・分類済みの大量の特許文書を教師文書データとしてその分類固有のタームを抽出する ・各分類の網羅範囲を規定した分類定義文章から分類に関係の深いタームを抽出することにより教師文書データに含まれないタームを補完する
4	発明内容と技術分野の類似性をどのように判定するか？	・付与対象特許文書からのターム集合と分類を特徴付けるターム集合間の類似度を算出する ・分類体系の階層に着目し，上位分類の類似度を下位分類の類似度算出に反映させる
5	分類体系更新後の分類を過去の大量特許にどのように付与し直すか？	（本研究の対象範囲外とする）
6	出力結果に対する利用者情報をどのようにフィードバックするか？	（本研究の対象範囲外とする）

［課題 4］発明内容と技術分野の類似性をどのように判定するか？

一般の分類自動付与において分類を特定する方式は大きく二つある．一つは KNN 法（K-Nearest Neighbors 法）に代表される方式である．すなわち，入力文章に類似する分類付与済み文書を検索し，検索結果上位 K 件の文書に付与されている分類の付与傾向を統計的に解析して付与すべき分類を特定する方法である．もう一つは，入力文章と各分類との間の類似度を算出し，高い類似度を持つ分類を付与する方法である．

分類付与に係る上記 4 課題に対して，それぞれどのような解決アプローチを採用するかを決定するにあたっては，図 1.5 に示した分類付与作業を前提とした特許分類自動付与に係る以下の三つの具備要件を考慮する必要がある．

［具備要件 1］上位分類の自動付与では，分類をノイズなく特定できること（正確性）
［具備要件 2］詳細分類の自動付与では，付与結果の根拠を提示できること（透明性）

［具備要件 3］分類付与に係るデータやプログラムがカスタマイズ可能なこと（拡張性）

以下，各具備要件について詳細に説明する．

［具備要件 1］上位分類の自動付与では，分類をノイズなく特定できること（正確性）

上位分類を自動付与することによって，特許文書を詳細分類付与担当者に自動的に振り分ける計算機システムを想定する場合，付与すべき分類（正解分類）が自動付与結果のできるだけ上位に少なくとも一つは含まれている必要がある．というのは，正解分類が上位に一つも含まれない場合，その特許文書は発明内容に全く関係のない技術分野の担当者に振り分けられてしまうため，文書の差し戻しが何度も発生して作業効率が悪くなるからである．しかし，自動付与結果の上位に正解分類を一つでも出力できれば，この特許文書を然るべき担当者の一人に早く振り分けることができる．振り分けることさえできれば，その担当者が自分の担当以外の技術分野の分類を付与すべきかを判断し，必要に応じてその技術分野の分類付与担当者に文書を回覧して分類付与を依頼することが運用レベルで可能となる．

［具備要件 2］詳細分類の自動付与では，付与結果の根拠を提示できること（透明性）

詳細分類付与作業において，計算機による自動付与結果を人手でチェックする作業プロセスを導入することを想定する場合，なぜその特許文書にその分類を付与すべきと計算機が判断したかを示す根拠を，担当者が容易に理解できる形で提示する機能が必要となる．例えば，その分類を付与するのに貢献したタームを提示したり，付与するのに貢献した入力文書中の記載箇所を強調表示したりする機能が考えられる．

また，正解分類をシステムが自動付与できなかった場合でも，担当者とのインタラクションによって，付与漏れした正解分類を容易に発見できるようにするための支援機能が必要である．例えば，自動付与された分類に意味的関連の深い分類群を提示することが考えられる．これらの支援機能が充実していないと，提示された分類付与候補のどれが妥当かを担当者が選択できないために自動付与結果を有効活用できず，担当者が結局始めから手作業で分類を付与することになりかねない．

［具備要件 3］分類付与に係るデータやプログラムがカスタマイズ可能なこと（拡張性）

一般に，分類を自動付与するためには，各分類の網羅範囲に関するデータや，各文書に記載される発明内容を認識するための用語辞書など，分類に係る知識データ（分類知識）が必要となる．技術トレンドや発明内容は時々刻々と変化するため，分類体系は更新されていくが，これに伴い分類知識も常に最新のものに更新していく必要がある．特許分類体

系が大規模であることを考慮すると，この更新作業は極力人手をかけずに簡単に遂行できることが望ましい．しかし，人手をかけずに分類知識を生成・更新すると，その品質に対する信頼性が低くなる．

そこで，日々の分類付与業務の中で，分類毎または分類付与担当者毎に分類知識を少しずつカスタマイズしていくことによって，分類知識をより高精度でより使いやすい形に最適化できる（使えば使うほど賢くなる）という拡張性が要求される．したがって，分類知識はこの最適化を可能とするようなデータ構造になっている必要がある．

上記と同様のことは分類自動付与プログラムの拡張にもあてはまる．分類自動付与プログラムが技術分野毎に異なっているとその保守コストが増大するため，分類自動付与プログラムのコアとなる部分は技術分野に依存しないことが望ましい．しかし一方で，各技術分野の特性に応じた分類付与アルゴリズムのカスタマイズを可能にすることも必要である．例えば，類似度を算出する式を技術分野毎にチューニングしたり，処理の中で使用しているパラメータ値を技術分野の特性に応じて最適化したりすることなどである．

上記三つの具備要件を踏まえ，上述した 4 課題のそれぞれに対する本研究での解決アプローチ（表 1.2）について詳細に説明する．

［課題 1 に対する本研究での解決アプローチ］

「特許文書に記載される発明内容をどのように記述するか」という課題 1 に対して，中間表現に変換するアプローチは，単語の意味属性や構文情報，文の意味を考慮するため，高い解析精度が期待できる．しかし，この解析には多くの語彙知識，言語固有の知識，技術分野固有の知識を必要とするため，これらの知識データを技術分野毎に構築・保守するには多大なコストがかかることが予想される．

一方，発明内容をタームの集合として記述するアプローチでは，単語や文の意味を考慮しないため，人間の直感と一致しない解析結果となることもしばしば起こる．しかし，用語辞書など最低限の知識があれば実現できるため，知識の構築及び保守が比較的低コストで済む．特許のように大規模な分類の付与を想定した場合，発明内容をタームの集合として記述するアプローチを採用する方が，具備要件 3（拡張性）で述べたデータのカスタマイズの容易性の観点から適切であると考える．

そこで本研究では，発明内容を重み付きタームの集合として表現する．すなわち，特許文書の発明内容を端的に表すタームを抽出して重要度に比例する重みを付与することにより，発明内容を記述するアプローチを採用する．

［課題 2 に対する本研究での解決アプローチ］

「特許文書に記載される発明内容をどのように特定するか」という課題 2 に対してであるが，特許文書は，新聞記事や科学技術論文などと比べると記述方法に癖のある文書である．1.3.1 節で述べたように，特許文書は特許タグからなる半構造化文書であり，どのタグにどんな内容を記述すべきかがある程度規定されている．

そこで本研究では，特許文書の構成（タグ情報）及び記載文章の構造（構文的特徴）に着目して発明内容を端的に表すタームを抽出する．また，特許分類体系の上位分類の多くが，発明に係る「もの」またはものに対する「処理」に基づいて構築されていることから，発明の構成要素に関する記載箇所よりも，その発明が適用される対象物または技術分野について端的に記載された箇所を重要視してターム抽出及び重み付けを行う．

これらのアプローチを具備要件 2（透明性）の観点で見ると，分類自動付与結果の根拠を提示する際に，単にその分類の付与に貢献したタームを提示するよりも，そのタームが発明の対象物を表すタームなのか，発明が適用される技術分野を表すタームなのかという情報も含めて提示できるので，分類付与担当者が分類自動付与結果の妥当性を判定することがより容易になると考える．

［課題 3 に対する本研究での解決アプローチ］

「各分類が網羅する技術分野範囲をどのように定式化するか」という課題 3 に対してであるが，特許分類体系を構成する各分類が網羅する範囲は，その技術分野の専門家が分類付与マニュアルとして文章で規定している．この文章の中には，各分類の網羅範囲を特定する適切なタームが多く含まれていると考えられる．一方で，過去に人手によって分類付与された特許文書が大量に蓄積されている．これらの特許文書を分類毎に解析することにより，各分類を特徴付けるタームを自動抽出できる．

分類の網羅する技術分野を特定するタームを抽出するという観点から見ると，上記分類付与マニュアル及び分類付与済みの特許文書は，リソースとしてそれぞれ一長一短がある．分類付与済みの特許文書から分類を特徴付けるタームを抽出する方法は，いわゆる事例ベース的なアプローチであり，比較的高精度なターム抽出が期待できる．しかし，その分類の技術分野を表すタームを漏れなく抽出できる保証はないため，タームの抽出漏れをなくすためには分類付与済みの特許文書データを大量に用意する必要がある．一方，分類付与マニュアルは，その分類の技術分野を文章で網羅的に規定しているので，技術分野の記載漏れはないが，そこで使用されるタームは特許文書で使用されるタームに比べて抽象度が高いことがしばしばある．その結果，分類付与マニュアル文章から抽出されたタームが，

分類付与対象となる特許文書から抽出されたタームと照合しないことが多い.

そこで本研究では，両者から抽出されたタームを統合することによって両者の欠点を相互補完し，より高精度な分類知識を生成するアプローチを採用する．すなわち，過去に分類付与済みの大量の特許文書を分類毎に収集し，その分類固有のタームを抽出する．その一方で，各分類の網羅範囲を規定した分類付与マニュアルからその分類に関係の深いタームを抽出し，分類付与済みの特許文書に含まれないタームを補完する．

［課題 4 に対する本研究での解決アプローチ］

「発明内容と技術分野の類似性をどのように判定するか」という課題 4 に対して，前述した KNN 法は，分類毎にタームを抽出する方式に比べて，以下の長所と短所を兼ね備える方式である．

［長所］ ・互いに内容が類似している文書が多い場合に有効である
・一つの文書に複数の分類を付与する場合の再現率が比較的良い
・文書検索と同じタームインデクスを用いることができる

［短所］ ・分類付与精度を更に向上させるためのデータカスタマイズが困難
・表記揺れによるターム照合漏れの影響が比較的大きい
・その分類を付与すべきと出力した根拠を提示しづらい

現状の分類付与技術では分類を高精度に自動付与することは難しいので，精度向上のためには人手によるカスタマイズが必要である．したがって，具備要件 3（拡張性）の観点から，分類知識や分類付与アルゴリズムを修正，チューニングできることが不可欠である．KNN 法は文書間の類似性から分類を推定する方法であるため，分類自動付与精度を向上させるためのチューニングは，文書単位で行わざるを得ない．したがって，比較的少数である分類毎にタームを抽出する方式に比べて，分類知識の拡張性の点で劣っていると考える．

そこで本研究では，各分類を特徴付けるターム集合を分類毎に抽出し，分類付与対象の特許文書から得られるターム集合と各分類を特徴付けるターム集合の間の類似度を算出するアプローチを採用する．

なお，上述以外の基本課題として，分類体系が更新される際に，過去に分類付与された大量の特許文書に対して更新後の分類をどのようにして付与し直すか，また，分類付与結果を出力した時にユーザから得られるフィードバック情報を，どのように分類付与方式に反映させるかなどが挙げられるが，これらは本研究の範囲外として本論文では扱わない．

1.4.2　類似特許文書検索研究の基本課題と解決アプローチ

一般に，文書検索精度を評価する際には，漏れとノイズの概念が導入される．「漏れがない」とは，検索結果の上位 N 件の中に所望の情報がすべて含まれていることを指し，「ノイズがない」とは，検索結果上位 N 件がすべて所望の情報であることを指す．漏れとノイズは一般にトレードオフの関係にあり，どちらかを改善しようとすると他の一方が悪化する．したがって，文書検索システムをどういう目的で使用するかによって，漏れ防止を重視したシステムとするか，ノイズ防止を重視したシステムとするかを決定する必要がある．例えば，特許庁における無効化文献の検索では，検索漏れの防止が絶対条件となるが，企業での技術動向分析において，ある特定の技術に係る特許を絞り込んで分析する場合は，分析精度を上げるためにノイズ防止も重視される．

本研究における類似特許文書検索の精度向上に係る研究の基本課題と，本研究におけるその解決アプローチを**表 1.3** に示す．本研究では，漏れの防止とノイズの防止という両面を考慮した検索アプローチを模索する．すなわち，検索におけるあるフェーズでは漏れ防止を意識し，別のフェーズではノイズ防止を意識した検索方式を検討する．

［課題 1］特許文書に記載される発明内容をどのように記述するか？

本課題は，1.4.1 節で述べた特許分類付与研究における課題 1 と共通である．すなわち，発明内容を記述する一般的な方式として，中間表現に変換する方式とターム集合として表現する方式があるが，類似特許文書検索の精度向上の観点から見た場合，どちらのアプローチを採用すべきかについて検討する必要がある．

［課題 1 に対する本研究での解決アプローチ］

本研究では，1.4.1 節の課題 1 に対する解決アプローチと同様，データのカスタマイズの容易性の観点から，発明内容を重み付きタームの集合として記述する．すなわち，特許文書の発明内容を端的に表すタームを抽出して重要度に比例する重みを付与することにより，発明内容を記述する．

［課題 2］特許文書に記載される発明内容をどのように特定するか？

本課題も，1.4.1 節で述べた特許分類付与研究における課題 2 と共通である．ただし，上位分類自動付与では技術分野に係るタームを重視するのに対して，特許検索では発明の新規性と進歩性（1.3.1 節参照）を端的に表すタームを重視するという大きな違いがあることを考慮しなければならない．

表 1.3　類似特許文書検索研究の基本課題と解決アプローチ

#	基本課題	本研究での解決アプローチ
1	特許文書に記載される発明内容をどのように記述するか？	・発明内容を重み付きタームの集合として記述する
2	特許文書に記載される発明内容をどのように特定するか？	・特許文書の構成及び請求項の文章構造に着目してタームを抽出し，重みを付与する ・発明の技術分野に係るタームよりも，発明の新規性・進歩性を表すタームを重視する
3	入力となる発明内容と特許文書の間の類似度をどのように算出するか？	・重み付きターム集合間の類似度を算出する ・類似する特許文書を段階的に絞り込む ・異なる観点で算出された類似度を統合する
4	出力結果に対する利用者からの情報をどのようにフィードバックするか？	（本研究の対象範囲外とする）

［課題 2 に対する本研究での解決アプローチ］

本研究では，1.4.1 節の課題 2 に対する解決アプローチと同様，特許文書の構成及び記載文章の構造に着目してタームを抽出し，重みを付与する．ただし，特許分類付与では発明に係る「もの」またはものに対する「処理」を表すタームに着目するのに対し，類似特許文書検索では発明の特徴を端的に記述した請求項に着目し，請求項の文章構造（図 1.3）を活用してターム抽出・重み付けを行うことにより，発明の内容をより正確に記述するというアプローチを採用する．

［課題 3］入力となる発明内容と特許文書の間の類似度をどのように算出するか？

類似度の算出方式は，検索の漏れ防止を重視するか，ノイズ防止を重視するかによって大きくアプローチが変わる．類似特許文書検索ではこれら両方に対する要求が高い．このことは，検索漏れ防止よりもノイズ防止が重要視される Web 検索とは異なる．

［課題 3 に対する本研究での解決アプローチ］

本研究では上記要求を鑑み，検索結果を漏れ防止とノイズ防止の両方の観点から段階的に絞り込むアプローチを採用する．検索漏れを防止するためには，いわゆる「広く浅い検索」が有効となる．すなわち，入力文章から得られる情報に少しでも関連しそうな特許文書は一次検索結果として残しておき，明らかに関連のない特許文書のみを一次検索結果から除外する．しかし，このままでは大量のノイズが含まれてしまうので，次に「狭く深いピンポイント検索」によって一次検索結果を並べ替える．すなわち，入力文章から発明の

特徴を端的に記述しているタームだけを抽出し，一次検索結果に含まれる特許文書の中から，それらのターム集合との類似度の高い文書を検索結果の上位に移動させることで，検索結果上位におけるノイズの発生を防止する．

なお，上述以外の基本課題として，出力検索結果に対する利用者からの情報をどのようにフィードバックするかという課題が挙げられるが，これについては本研究の範囲外として本論文では扱わない．

1.5　本論文の構成

本研究は，特許文書の記載方法に関する特徴を踏まえ，特許文書の構成及び特許文書に記載される文の構文的・語彙的特徴を利用したターム抽出・重み付け方式及び文書間（または文書と分類の間）の類似度算出方式を提案している点が従来研究と異なる点である．

以下，本論文の構成について述べる．

第 2 章では，特許分類体系の上位分類である「テーマ」及びテーマのさらなる上位分類である「審査室」を特許文書に自動付与する技術について述べる．ここでは特許文書の構成及び文章構造に着目し，タームの出現箇所及び出現共起性を手掛かりとして発明に係る対象物及び技術分野を端的に記述するタームを抽出し，その重要度に比例する重みを付与する方式を提案している．また，教師文書データとなる異種の文書（過去に付与済みの特許文書及び分類付与マニュアル）から各分類を特徴付けるタームを特定する方式と，付与対象となる特許文書から抽出されたタームとの照合によって付与すべき分類を特定する方式を提案している．更に，分類体系の階層に着目して分類を特定する二段階分類付与方式を提案している．これらの方式により，テーマ 2,815 分類の中から約 62%の正解率で適切なテーマを自動付与することができるという実験結果を得ている．

第 3 章では，ある特許文書に類似する過去の特許文書を高精度に検索する類似特許文書検索技術について述べる．ここでは，発明内容を端的に記載した請求項文章を入力として，そこに記載された発明内容に類似する特許文書を検索する．その際に，特許文書の構造，特に請求項の文章構造に着目して検索タームを抽出，重み付けし，特許文書間の類似度を算出する方法として二段階検索方式を提案している．第一段階では再現率重視の「広く浅い」検索手法を採用することにより，できるだけ多くの類似特許文書が検索結果に含まれるようにする．第二段階では適合率重視の「狭く深い」検索手法を採用することにより，真に類似する特許文書が検索結果の上位にランクされるように，第一段階で得られた検索結果集合を並べ替える．また，請求項固有の記述特性を利用したタームの重み付け方式も

併せて提案している．これらの方式により，評価に用いたデータセットによって傾向のばらつきはあるものの，全体として検索精度が向上することを確認している．

第 4 章では，第 3 章の類似特許文書検索技術に関連して，出願特許の出願人（特許の権利化を申請する組織）と，その発明を無効化する特許の出願人との同一性が，類似特許文書検索精度に与える影響について考察する．まず，出願特許とその無効化特許の出願人に関する傾向を，文書属性，使用タームの共通性，検索の難易度という三つの観点から定量的に分析している．分析結果から，(1)出願特許とその無効化特許の出願人が同じとなる現象は，無効化特許件数の約 21%で起きており，どの技術分野や出願人にも見られる一般的な現象であること，(2)出願特許とその無効化特許の出願人が同じ場合，共通して使われるタームの割合が比較的高いこと，(3)出願人が出願特許と違う無効化特許の検索は，出願人が同じ場合に比べて困難なこと，(4)出願特許とその無効化特許の出願人が同じか違うかによって，適用する検索方式の精度的振る舞いが変わること，といった知見を得ている．次に，これらの知見に基づいて，出願特許とその無効化特許の出願人の同一性に着目した検索手法を提案している．本手法は，出願人の同一性の観点から個々の検索方式の有効性を評価し，その結果を踏まえて，複数の検索方式を適切に組み合わせることを特徴とする．そして最後に，類似特許文書検索方式を検討・評価する際には，入力特許と検索されるべき正解特許との間の出願人の同一性を考慮すべきであると結論付けている．

第 5 章では，特許文書を含めた分類付与技術及び文書検索技術に係る従来研究動向について述べ，本研究と比較する．まず，文書検索の歴史的経緯について簡単に触れる．次に，文書検索に係る研究動向として，特に検索モデル，インデクシング方式，ターム抽出及び重み付け方式，検索アルゴリズムを採り上げ，それぞれ簡潔に説明する．次に，分類自動付与に係る研究動向について述べる．最後に，特許文書を対象とした類似特許文書検索及び特許分類自動付与に係る研究動向について述べ，本研究のアプローチと比較する．

第 6 章では，上記の研究成果として得られた知見を総括するとともに，今後の研究課題について述べる．また，本研究成果の特許以外の文書への拡張性についても言及する．今後は精度向上だけでなく，作業担当者と計算機の役割分担を明確にして，作業を協調的に行える支援環境を併せて考えていくことが重要になってくると考えている．

第２章　文書内の言語構造を利用した特許分類自動付与

2.1　はじめに

特許庁において出願特許を審査する際には，出願特許文書に記載された発明内容を無効化する特許（無効化特許）が過去に出願されていないかを検索調査する．しかし，検索対象となる特許文書は数百万件にのぼり，技術分野も多岐に渡っており，また検索漏れが絶対に許されないことから，無効化特許の検索には多大なコストがかかる．

そこで，特許庁では以前より，出願特許文書に対してその発明の技術分野・特徴に基づいて種々の分類を付与[5]している[6]．無効化特許の検索においてこの分類情報を活用することにより，検索対象となる特許文書集合を大幅に絞り込むことができる．また，分類情報は特許の出願動向分析にも不可欠な属性である．

しかし，特許の出願件数が年間 40 万件を超えており，1 日あたり約 2,000 件の特許文書に対して分類を付与する必要がある．しかし，分類体系が大規模で複雑なためにその技術分野に精通した分類付与の専門家が手作業で分類を付与せざるを得ず，毎年多大な作業時間と人件費を費やしているのが現状である．特許審査期間の一層の短縮が叫ばれている現在，計算機による特許分類自動付与への要求は高まる一方である．

人間が文書に分類を付与する際には，各分類がどこからどこまでの技術分野を網羅しているのかを把握していなければならない．このことは計算機による分類自動付与においても言えることであり，各分類の網羅範囲を規定する分類知識が不可欠である．しかし，特許が網羅する技術分野は広範囲に渡っており，すべての技術分野に対してそれぞれ高精度の分類知識を構築，保守するためには莫大なコストがかかることが予想される．

そこで本研究では，この分類知識を自動生成し，特許文書を入力としてその発明内容に係る技術分野を特定し，分類知識を参照して適切な分類を自動付与する方式を提案する[59][60]．本方式では，各分類に属する特許文書群から抽出した重み付きタームの集合によってその分類が網羅する技術分野を定式化するアプローチを採用する．そして，分類付与対象の特許文書に記載されるターム集合との類似度を算出し，類似度の高い分類をその

[5] 本論文では，分類体系を構成する要素を「分類」と記述し，一般に使われている「分類する」という動作を「分類を付与する」と記述することとする．

[6] 実際には，特許庁の外郭団体である（財）工業所有権協力センターが分類付与作業を請け負っている．

特許文書に付与する．

本研究で提案する特許分類自動付与方式は,以下の4種類の処理方式を特徴としている．

(1) 特許文書の言語構造に着目し，特定の特許タグのみから発明の技術分野を特定するタームを抽出する方式（2.2.1節）

(2) 特許文書におけるタームの出現位置及び出現共起に基づいてタームに重みを付与する方式（2.2.2節）

(3) 上記ターム抽出及び重み付け方式を用いて分類知識を自動生成する方式として，大量の分類付与済み特許文書と，各分類の適用範囲を文章で規定した分類付与マニュアルからそれぞれ抽出した重み付きタームを統合し，各分類を特徴付ける重み付きターム集合からなる分類知識を自動生成する方式（2.3節）．

(4) タームの分類別出現傾向及び分類体系の階層性に着目して，文書と分類との間の類似度を算出する方式（2.4節）

本研究で対象とする特許分類体系として，F タームの上位分類であるテーマ（2,815 分類）と，テーマの上位分類である審査室（38 分類）を対象とする．これらを付与対象として選定した理由として以下が挙げられる．

(1) 分類付与担当者はテーマ単位で割り当てられているため，テーマが確定すると分類付与担当者を割り振ることができる．担当者は割り当てられた特許文書に対して更に詳細な分類（FI やF タームなど）を付与する．

(2) テーマと FI の間は 1 対多の対応関係があり，また，F タームはテーマの詳細分類であることから，テーマが確定すると，FI とF タームを並列作業で付与することが理論的に可能となる．その際，テーマが既に確定しているため付与すべき FI 及びF タームは少数に絞られることから,分類付与精度の向上及び分類付与作業の短縮が期待できる．

(3) テーマの粒度（2,815 分類）は計算機によって自動付与できる現実的な粒度である．

以下，2.2節，2.3節，2.4節では，本研究で提案する特許分類自動付与方式の特徴となる上記 4 種類の処理方式について詳細に述べる．2.5節では，これらの処理方式に基づく分類自動付与の処理手順について述べる．2.6節では，最大 31 万件の公開特許公報データを用いた分類自動付与精度の評価実験結果を示し，それについて考察する．2.7節では，特許分類自動付与研究の成果についてまとめる．

2.2　文章解析方式

2.2.1　文書内の言語構造を利用したターム抽出方式

一般に，分類の網羅範囲を規定するターム集合は，分類の粒度によってその抽出方法が異なる．分類の粒度が粗い場合，例えば，「工作機械」「情報処理」「医療」のような，発明の具体的内容でなく発明の適用される技術分野を特定する一般的/抽象的なタームが有用な手掛かりとなる．逆に分類の粒度が細かくなると，分類の網羅範囲の差異を表すタームは，例えば「歯車加工」「文書検索」「磁気治療器」など，より専門的/具体的なタームになってくる．

さて，新聞記事が世の中の「出来事」を記述した文章であるのに対し，特許文書は発明に係る「もの」またはものに対する「処理」を記述した文章であると言える．各テーマの網羅範囲を記述した分類付与マニュアルを見ても，「もの」または「処理」に基づいて網羅範囲を規定しているテーマが多いことが分かる．したがって，特許文書に記載される「もの」あるいは「処理」を端的に表すタームが抽出できれば，対応するテーマを特定できると考えられる．

本研究では，発明に係る「もの」あるいは「処理」を表すタームを特定するために，特許文書を構成するタグに着目する．すなわち，上記タームを含む可能性が高いタグとして，以下の三つのタグを採り上げる．

(1) 発明内容を端的に記述した【発明の名称】タグ
(2) 発明の構成要素または処理ステップ，発明の適用対象を記述した【請求項】タグ
(3) その発明がどの技術分野で適用されるかが具体的に記述された【技術分野】タグ

上記三つのタグのうち，【請求項】タグ及び【技術分野】タグに記述される文は，後述するように特定の構文を有することが多い．これらの構文情報を利用することにより，技術分野を特定するのに有用なタームを抽出できると考える．

また上記タグに対し，【発明の実施の形態】タグでは，発明を実現する方法が詳細かつ具体的に記述される．この中にも技術分野を特定するのに有用なタームが多く含まれている反面，技術分野に関係のないノイズタームも多く含まれている．このタグに記載される文章は文章量が多い上に，タームを抽出するのに有効な構文的特徴がほとんど見られないので，技術分野を特定するタームだけを高精度に抽出することは難しいと考える．

また，【要約】タグは，発明内容を 400 字以内でまとめたものであり，技術分野を特定するのに有用なタームも含まれている．しかし，【発明の実施の形態】タグと同様に構文

的特徴が見られないこと，記載内容自体が適切でない場合が多いこと，使用されているタームが【発明の名称】,【請求項】,【技術分野】の 3 タグで使用されているタームと重複していることが多いことから，技術分野の特定に有用なタームを抽出しづらいと考える（500 件の特許文書による分析調査では，【要約】タグ中のタームの約 69.0%が上記 3 タグにも出現した）.

各タグにおける文章記載に係る上記特徴を踏まえ，本研究では【発明の名称】,【請求項】,【技術分野】の 3 タグのみをターム抽出の解析対象とする．ただし，【技術分野】タグについてはノイズターム除去の観点から，「本発明」「本願」「この発明」「発明は」の 4 表現のいずれかで始まる最初の一文のみを解析対象とする（500 件の特許文書による分析調査では，99.4%が上記 4 表現のいずれかで始まる文を含んでいた）.

また，本方式では，「もの」あるいは「処理」を端的に表し，技術分野の特定に有用なタームとしてその品詞に着目する．すなわち，名詞，動詞語幹（サ変動詞含む），単位語（「mol」「Hz」など技術分野固有のものが多い），アルファベット列，カタカナ列，未知語のみをタームとして抽出する．

2.2.2 タームの出現位置と出現共起を利用したターム重み付け方式

ターム抽出の解析対象となる【発明の名称】,【請求項】,【技術分野】の 3 タグから抽出されたタームへの重み付け方式として，本方式ではタームの出現位置及びタームの出現共起という 2 種類の手掛かりに着目する．

(1) タームの出現位置を利用したターム重み付け

ターム抽出の解析対象となる上記 3 タグのうち，【請求項】タグ及び【技術分野】タグについては，記述される文の構文が比較的定型である．【請求項】タグでは次の定型文で記述される場合が多い（500 件の特許文書による分析調査では，71.4%について下記構文を満たしていた）.

『[～において,] ～A を特徴とする B』

また，【技術分野】タグでは，次の構文及びこれに類する構文で記述されることがほとんどである（500 件の特許文書による分析調査では，97.6%について，下記構文を満たしていた）.

『本（この）発明は，（C に関し,）～D に関する』

上記二つの定型文の A，B，C，D の記述部分に含まれるタームは，発明の内容または技術

表 2.1　ターム出現位置と出現共起に基づくターム重み設定ルール

#	ターム重み配分条件	重み
1	【請求項】の文末が「ＡのＢ」なる名詞句である場合のタームＡ	5
2	【請求項】の文末の名詞（句）で，#1 以外のターム	3
3	【請求項】において「Ａを特徴とするＸ」のＡに相当する名詞句を構成するターム	1
4	【請求項】に出現する単語	2
5	【発明の名称】に出現する単語	3
6	【技術分野】において「本発明」「本願」「この発明」「発明は」のどれかで始まる文において「Ａに関する」「Ａに関した」「Ａに関し」「Ａに関して」「Ａについて」「Ａのための」「Ａのために」「Ａにおける」のどれかの構文を満たす名詞句Ａを構成するターム	3
7	【技術分野】において「本発明」「本願」「この発明」「発明は」で始まる文に出現するターム	2
8	#1 または#2，#5，#6 の 3 条件をすべて満たすターム	10
9	#1 または#2，#5，#6 の 3 条件のうち，2 条件のみを満たすターム	7
10	#4 及び#7 を満たす単語	4

注 1：複数の条件を満たす場合，各重みの値の合計値がそのタームの重みとなる．
注 2：2.6 節の評価実験で用いた分類知識を生成する際には，重みが 2 以下のタームを除外した．

分野を特定または限定する重要なタームとみなすことができる．そこで，これらの位置に出現するタームの重みを他の位置に出現するタームの重みよりも大きくする．

(2) タームの出現共起情報を用いたターム重み付け

本方式では，「技術分野の特定に有用なタームは，解析対象となる上記 3 タグのうちの複数に出現する」と仮定する．そこで，あるタームが複数のタグに共起して出現する場合，そのタームの重みをそうでないタームよりも大きくする．

タームの出現位置と出現共起に基づいたタームの重み設定ルールを**表 2.1** に示す．表中の#1～#7 がタームの出現位置に基づく重み設定ルールであり，#8～#10 がタームの出現共起に基づく重み設定ルールである．タームの出現共起の度合が高いほど，タームの重みが大きくなるように設定している．なお，これらの重みの設定については，分類付与精度を比較しながら重みの値を試行錯誤的にチューニングし，最も付与精度が良い重みを採用している．

2.3　教師文書データからの分類知識自動生成方式

各分類の網羅範囲を規定し，それを分類知識として定式化するのは人間にとっても難しく，またそれが可能だとしても多大な労力を必要とする．本研究では，各分類の網羅範囲を緻密に規定する代わりに，どんなタームがどのくらいの重要度（重み）で各分類を特徴付けているかを規定する．そして，分類別のタームとその重要度を分類知識として記述す

る．すなわち分類知識は，分類とその分類を特徴付けるターム及びその重要度（重み）の3要素からなるレコードで構成される．以下に分類知識の例を示す．

（フォーマット）『分類名称，ターム文字列，重み』
（データ記述例）通信　　，伝送　　，20
通信　　，ネットワーク，50
計算機応用，ネットワーク，20
……

ここで，ターム「伝送」は分類「通信」を特徴付けるタームで，その重要度（重み）が20であることを示し，ターム「ネットワーク」は分類「通信」及び「計算機応用」の両方を特徴付けるタームで，その重要度（重み）はそれぞれ50，20であることを示している．

本方式では，上記の構造を持つ分類知識を自動生成する．各分類を特徴付けるタームを収集するために，教師文書データを用いる．一般に，教師文書データからの知識獲得では，教師文書データの選定方法を考慮する必要がある．本方式では，分類知識を生成する際に使用する教師文書データとして，既に分類付与済みの特許文書と，各分類の網羅範囲を人間が文章で規定した分類付与マニュアルの2種類を用いる．特許のように技術分野が広範囲で分類が大規模の場合，分類の適用範囲をすべて網羅するのに十分な分類付与済み特許文書を選別・収集することは困難である．一方，分類付与マニュアルは，分類付与済み特許文書で網羅できなかった範囲に関するタームを分類知識に追加補足する役目を持ち，分類自動付与精度を更に向上できると考える．

ここで問題となるのは，分類付与済み特許文書と分類付与マニュアルという，異なる2種類の教師文書データから抽出されたタームを，分類付与時にどのように活用するかである．分類付与済み特許文書から抽出されたタームには，2.2.2節で述べたようにタームの出現位置及び出現共起に基づく重み付けがなされる．一方，分類付与マニュアルは特許文書のように文章が構造化されていないので，抽出されるタームの重み付けは出現頻度などを手がかりとせざるを得ない．したがって，このように異なる種類の文書から異なる観点に基づいて付与された重みの価値を統一して分類知識を生成する必要がある．

分類付与済み特許文書及び分類付与マニュアルの各々から抽出された分類別のターム数の重み分布を**図2.1**に示す．横軸の数値は，特許文書，分類付与マニュアルのそれぞれから抽出されたタームに付与された重みを，分類別かつターム別に合計した値である．分類付与マニュアルから抽出したタームの重みは出現頻度の値をそのまま用いている．分類付与済み特許文書から抽出したタームの重み分布と，分類付与マニュアルから抽出したタ

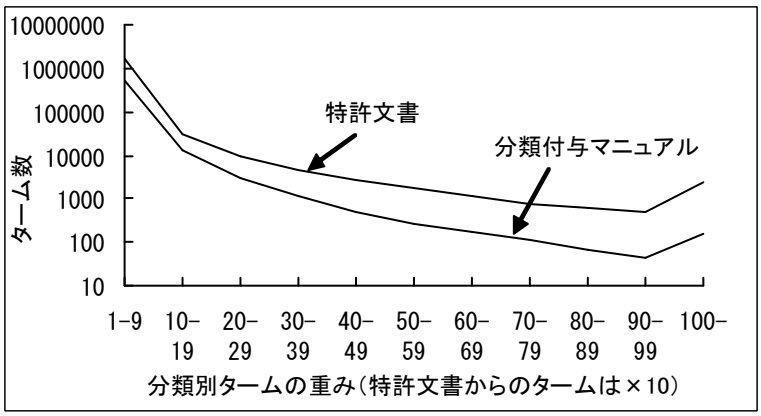


図 2.1　2 種類の教師文書データから抽出される分類別タームの重み分布

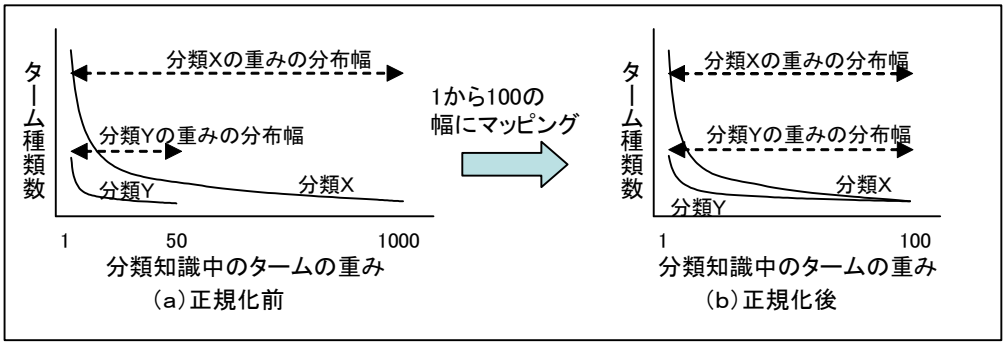


図 2.2　分類知識における分類別タームの重みの正規化方法

ームの重み分布は，重みの大きさは違うものの分布傾向は類似していることが分かる．

そこで本方式では，分類付与済み特許文書から抽出したタームの重みを単純にある定数 C で割ることによってそれぞれの重みの価値がほぼ同じになるように補正した後に，両者の重みを合算することにより，統合後のタームの重みを算出する．ただし定数 C の値は経験的に設定する必要がある．

また，統合後のタームの重みの分布範囲は，分類によってかなりばらつく．このばらつきはタームの重要度の違いによる以外に，分類別の教師文書データ量の偏りによるところが大きい．教師文書データから抽出されたタームを分類別にまとめる時にその重みを単純に合算しているため，教師データ量が多い分類のタームの重みは必要以上に大きくなってしまう．

そこで**図 2.2** に示すように，分類毎に重みの分布幅を一定にし（正規化），重みの価値が分類によらず同一になるように重みをマッピングすることにより，上記問題を解決する．正規化には偏差値の概念を導入する．すなわち各分類について，ある偏差値のしきい値に対応する重みの値を求め，その値が上限となるように重みを修正した後，各タームの重みを 1 から 100 の間にマッピングする．この時，最大値に対する相対比は保存される．この結果，すべての分類のすべてのタームの重みは 1 から 100 の間に再配分される（ある分類において正規化前に最小の重みを持っていたタームは重み 1 に，上限値以上の重みを持っていたタームはすべて重み 100 に正規化される）．

2.4 類似度算出方式

2.4.1 分類別の類似度算出方式

新規の特許文書に分類を付与する場合，分類知識生成時と同一の方法で特許文書からタームを抽出して重みを付与する．そして抽出されたターム集合と分類知識中のターム集合との照合を分類毎に行い，各分類に対する類似度を計算し，類似度の高い分類を出力する．

本方式では，次式により新規特許文書と各分類の間の類似度を計算する．

$$\left.\begin{aligned} S(i) &= \sum_{j=1}^{n} s(i,j) \\ s(i,j) &= W(j) \times \sqrt{\frac{w(i,j)}{\sum_{k=1}^{m} w(k,j)}} \end{aligned}\right\} \quad (2.1)$$

ここで，S(i)は新規特許文書と分類 i の類似度，s(i,j)は新規特許文書から抽出されたターム j に対する分類 i の類似度,n は新規特許文書から抽出されたタームの異なり数,W(j)は新規特許文書におけるターム j の重み，w(i,j)は分類知識中の分類 i におけるターム j の重み,w(k,j)は分類知識中の分類 k におけるターム j の重み,m は分類数を示している．本算出方式は次の性質を持つ．

(1) 新規特許文書におけるタームの重み W(j)が高いと類似度も高い．
(2) 分類知識中のタームの重み w(i,j)が高いと類似度も高い．
(3) 多くの分類に現れる（Σw(k,j)の値が大きい）タームは類似度が低い．

本方式では，文書検索におけるターム重み付けで良く使われる IDF（Inverted Document

Frequency) を使用していない．その代わりとして，タームの出現分類数を考慮している．すなわち，分類の特徴を表さない一般語は，技術分野をまたがって多くの分類に出現するため，出現分類数は大きくなる．そこで，このような場合の類似度は低くする（式(2.1)で，平方根内の値が小さくなる)．逆に，ある分類を特徴付けるタームは出現分類数が小さくなるので，このような場合の類似度は高くする．

また式(2.1)では，平方根演算子によってターム別の類似度を補正している．あるタームが分類知識の中に含まれている分類にはそのタームに係る類似度が加算されるが，この平方根がない場合，その類似度は式(2.1)の(w(i,j)/Σw(k,j))の値に比例する．しかし，直感的にはそのタームを含んでいる分類は重要であるので，そのタームを含んでいない分類との間の類似度の得点差は大きくすべきである．逆に，大変重要なタームであり，式(2.1)の(w(i,j)/Σw(k,j))の値が 1 に近い場合，どれも重要とみなされるのであるから，得点差は小さくしてもよいであろう．そこで，類似度計算において上記の性質を近似的に備えている演算子として平方根を採用し，類似度の値を補正している．

2.4.2　分類体系の階層に基づく二段階分類付与方式

特許分類のように大規模で階層構造をなす場合，分類体系の下位分類をいきなり特定するよりも，付与すべき分類をまず上位分類で少数に絞り込んだ後に下位分類を付与することにより，結果にノイズが含まれるのをある程度防ぐことができると考える．そこで，テーマの上位分類である審査室の分類付与結果を，下位分類であるテーマの付与に反映させる二段階分類付与方式を提案する．

まず，これまでに述べた手法により，新規特許文書に審査室分類を付与する．次に審査室の下位分類であるテーマを付与する際に，審査室付与で算出された各審査室の類似度の相対値を，各テーマの類似度計算結果に乗ずる．これにより，審査室付与で上位にランクされた審査室に属するテーマが分類自動付与結果の上位に出力されやすくする．すなわち，次の式により各テーマの類似度を算出する．

$$\left.\begin{aligned} S(i) &= \mathrm{Bonus}(i) \times \sum_{j=1}^{n} s(i,j) \\ \mathrm{Bonus}(i) &= \frac{Su(\mathrm{upper}(i))}{\sum_{k=1}^{m} Su(k)} \end{aligned}\right\} \quad (2.2)$$

ここで，S(i)は新規特許文書とテーマ i の類似度，Bonus(i)はテーマ i の類似度に乗ぜら

れる係数，s(i,j)は新規特許文書から抽出されたターム j に対するテーマ i の類似度（式(2.1)により算出），n は新規特許文書から抽出されたタームの異なり数，upper(i)はテーマ i が属する審査室，Su(k)は審査室 k の類似度，m は審査室数を示している．

別の方式として，審査室の付与結果から審査室を上位数個に絞り込み，他を足切りするという方式も考えられる．審査室の付与精度が 100%に近いほどこの方式は有効であると思われるが，そうでない場合，上位分類である審査室の付与を誤ると下位分類であるテーマ付与では必ず誤ってしまう．したがって，上位分類の付与精度に応じてどちらが良いかを選択するのが最善であると考える（2.6 節の評価実験では式(2.2)を採用している）．

2.5　特許分類自動付与の処理フロー

前節までで述べた分類自動付与方式の処理フローを**図 2.3** に示す．

(1) 分類知識の生成処理フロー

新規特許文書への分類付与に先立って，分類知識を自動生成する．本処理の入力は，分類付与済み特許文書及び分類付与マニュアル文章である．

図 2.3 において，解析対象タグ文章抽出では，分類付与済み特許文書から【発明の名称】，【請求項】，【技術分野】の各タグに書かれた文章を抽出する．ただし，【技術分野】については，2.2.1 節で述べた構文を満たす文のみを抽出する．

ターム抽出では，まず形態素解析によって文章を単語に分割する．ここでは，17 万語彙からなる単語辞書を参照して単語見出し情報と品詞情報を取得する．単語辞書にない単語は未知語とみなされる．そして，品詞情報に基づいて分類付与で用いるタームを抽出する．

後処理では，後のターム照合の精度を向上させるために，得られたタームについてカタカナ小文字の大文字化及び長音記載の統一化を行う（例：「チョコレート」⇒「チヨコレート」）．

不要語除去では，予め用意した 713 語の不要語（「発明」「特許」など）を除去する．

ターム重み付けでは，分類付与済み特許文書から得られたタームについては，その出現位置及び出現共起に基づいて，表 2.1 で示した重み設定ルールを適用することによって重みを付与する．分類付与マニュアル文章から得られたタームについては，その出現頻度をそのまま重みとして付与する．

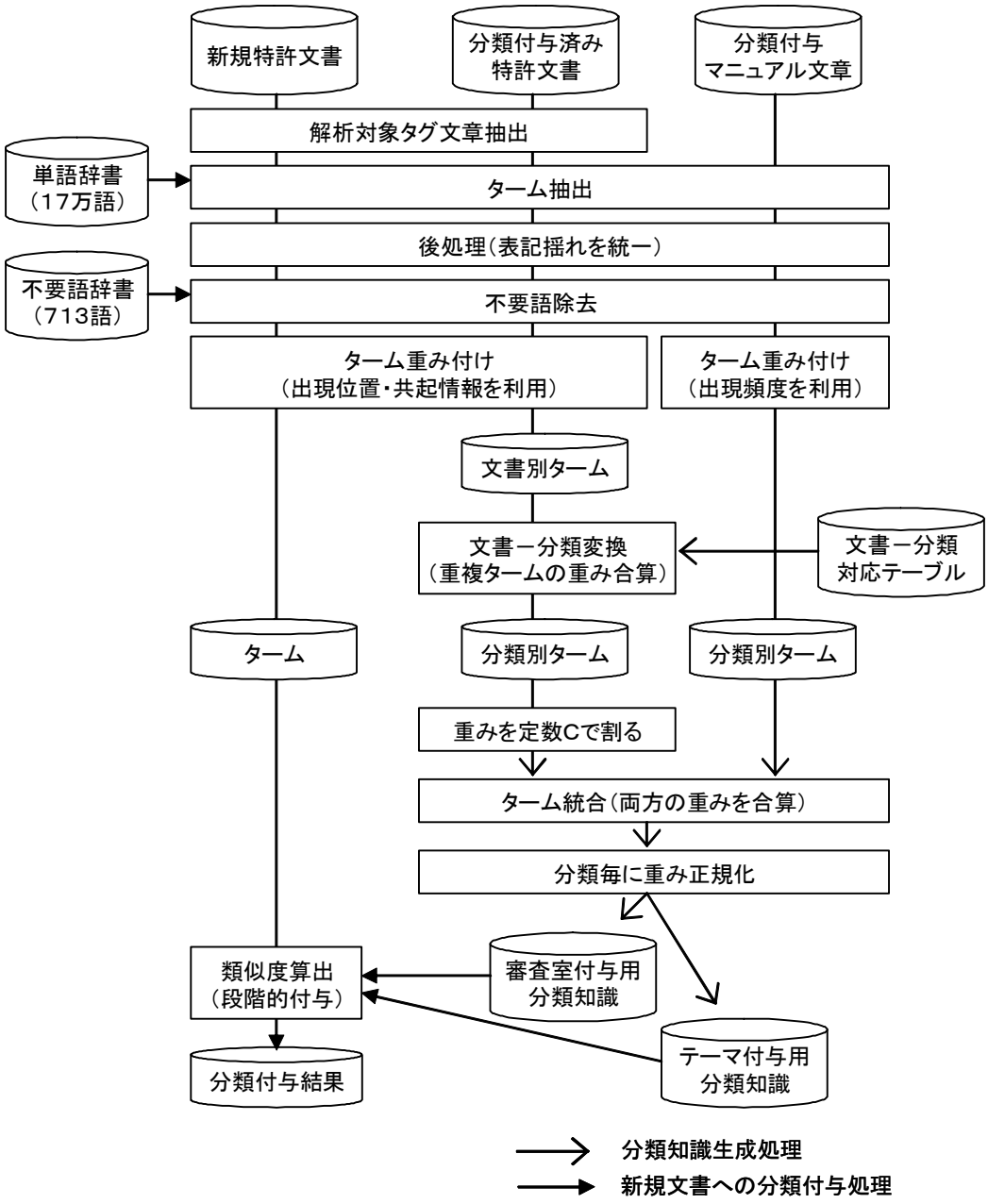


図 2.3　特許分類自動付与の処理フロー

文書－分類変換では，特許文書に付与されている分類に基づいて，文書別タームを分類別タームに変換する．この際に，同一の分類内で重複するタームについてはその重みを合算して一つにまとめる．そして，分類付与済み特許文書から得られたタームの重みを定数 C で割った後に，分類付与マニュアル文章から抽出されたタームと統合する．この際にも，

同一の分類内で重複するタームの重みを合算して一つにまとめる．そして最後に，図 2.2 に示した方法によって分類毎にタームの重みを正規化する．

上記処理を審査室単位及びテーマ単位で行い，それぞれに対応する分類知識を生成する．

(2) 新規特許文書への分類付与処理フロー

新規特許文書に分類を付与する処理は，ターム重み付けまでは分類知識生成時と同一である．類似度算出では，新規特許文書から抽出された重み付きタームと，分類知識に記述された分類別の重み付きタームを照合し，式(2.2)に示した算出式を用いて分類毎に類似度を算出する．そして，類似度の高い分類を自動付与結果として出力する．

2.6 分類自動付与精度評価実験

2.6.1 実験方法

テーマ（2,815 分類）及びその上位分類である審査室（38 分類）を対象とした分類自動付与精度の評価実験を行った．分類知識生成用の分類付与済み特許文書データとして，最大 31 万件の公開特許公報データを用いた．また評価用特許文書データとして，最大 16 万 6,000 件の公開特許公報データを用いた．各公開特許公報には分類付与の専門家によりテーマが平均 1.9 個，審査室が平均 1.5 個付与されている．

分類自動付与精度を表す尺度として，再現率と適合率（精度）が有名である．再現率が高いほど分類の付与漏れが少なく，適合率が高いほどノイズ分類が少ない．両者の間には一般にトレードオフの関係がある．しかし，本実験では再現率，適合率の代わりに以下の尺度を「正解率」として用いる．

$$[\text{上位N位の正解率}] = \frac{[\text{上位N位に正解を一つ以上含む評価文書件数}]}{[\text{評価文書件数}]} \qquad (2.3)$$

すなわち，ある評価用特許文書に対してシステムが自動付与結果として出力した上位 N 個の分類の中に，専門家が付与した分類（正解分類）が「少なくとも一つ以上」含まれている場合，その特許文書への分類自動付与結果は正解であるとし，正解した評価文書件数の割合(%)を正解率とする．この評価尺度は，1.4.1 節で述べた具備要件 1（正確性）に対応して新しく定義したものである．すなわち，分類付与すべきテーマのうちの一つでも自動付与できれば，F タームや FI などの詳細分類を付与する担当者を確定でき，全体としての分類付与作業効率を向上できるとの認識に基づいている．

表 2.2　分類自動付与精度評価実験の内容一覧

#	実験目的・内容	教師文書データ	評価データ
1	・教師文書データの種類の違いによる分類付与精度比較 ・必要十分な分類付与済み特許文書の量の把握	(1) 公開特許公報 1-31 万件 (2) 分類付与マニュアル文章 (3) 上記 2 種類を統合	公開特許公報 3,850 件
2	ターム重み付けに係るパラメータチューニングによる分類付与精度向上の検証	公開特許公報 15-31 万件と分類付与マニュアル文章を統合	
3	大量の評価データによる分類付与精度評価		公開特許公報 16 万 6,000 件
4	全自動分類付与システムとしての実現可能性検証		

本研究では，**表 2.2** に示す 4 種類の実験を行う．

【実験 1】教師特許文書データの種類及び規模の違いによる分類付与精度比較

分類付与済み特許文書及び分類付与マニュアル文章から自動生成した分類知識を用いて分類を自動付与した時の精度を比較検証する．分類知識生成の入力データとして，(1)分類付与マニュアル文章のみ，(2)分類付与済み特許文書のみ（1 万件から 31 万件まで増加させる),(3)両者を統合した場合,の3パターンについて分類自動付与精度を測定する．また，分類知識生成用の分類付与済み特許文書件数を増加させた時の分類自動付与精度の変化を検証することにより，分類知識生成に必要十分な分類付与済み特許文書の量を把握できる．これは分類知識保守の観点（分類付与済み特許文書の収集作業・解析時間の効率化）から重要なデータとなる．なお，評価用文書は，新規公開特許公報 3,850 件を用いる．

【実験 2】ターム重み付けに係るパラメータチューニングによる分類付与精度向上の検証

ターム重み付けに係る各種パラメータをチューニング（最適化）することによって分類自動付与精度を向上できると考える．重みのチューニング方法として，(1)分類知識生成において分類別のタームの重みを合算する際の合算値の上限値の設定，(2)公開特許公報中のタームと分類付与マニュアル文章中のターム統合方法の最適化（図 2.3 における定数 C の値の設定），(3)分類知識におけるターム重みの正規化適用の有無，の 3 種類についてその有効性を評価する（本節では各パラメータの最適値における分類自動付与精度のみについて述べる）．

【実験 3】大量の評価データによる分類付与精度評価

上記 2 種類の実験で用いる評価用文書データは 3,850 件であるが，2,815 分類という分類体系の規模を考慮すると量が絶対的に少ない．そこで本実験では約 16 万 6,000 件の公開特許公報を評価用文書データとすることにより，本方式が精度的に安定しているかを検

証する．分類知識は 2 種類の教師文書データを統合したものを用い，パラメータは実験 2 で最適化した後のものを用いる．

【実験 4】全自動分類付与システムとしての実現可能性検証

本実験では，全自動で漏れなく審査室を付与できるかという観点から分類自動付与精度を検証する．すなわち，実験 3 の審査室自動付与結果を別の評価尺度，すなわち自動付与された上位 3 個の審査室の中に，専門家が付与した正解審査室がいくつ含まれているかという再現率によって評価する．

2.6.2 実験結果と考察

【実験 1】教師文書データの種類及び規模の違いによる分類付与精度比較

今回の実験で用いた分類知識を構成するタームに関する統計データを**表 2.3** に示す．公開特許公報 15 万件から抽出されたターム種類数（異なり数）の方が，分類付与マニュアル文章から抽出されたターム種類数よりも多い．また，公開特許公報から一つもタームを収集できなかったテーマが 26 分類あった．

実験結果を**図 2.4** に示す．図 2.4 の横軸は，分類知識作成用の教師文書データの件数であり，1 万件から 15 万件までの 1 万件刻みと 31 万件で分類付与精度を算出している．縦軸は上位 3 位での正解率（式(2.3)で N=3）である．分類自動付与の正解率は 2 種類の教師文書データを統合した場合が最も高く，分類付与済み特許文書のみの場合に比べて 2.2 ポイント（76.4%⇒78.6%）から 10.5 ポイント（59.4%⇒69.9%）の範囲で向上している．これは，分類付与済み特許文書から抽出できなかったタームを分類付与マニュアル文章中のタームが補完していると考えられる．一方，分類付与マニュアル文章のみの場合の正解率は他の二つに比べて大きく低下した．この理由として，分類付与マニュアル文章のデータ量が比較的少ないこと，記述フォーマット（文章構造）が特許文書と異なること，分類付与マニュアル文章では分類の範囲を規定する抽象的/集合的なタームが使われるために特許文書で使われるタームと照合されにくいこと，などが挙げられる．したがって，分類付与マニュアル文章は補完的なデータとして位置付けるべきであろう．

また，教師文書データとして使用する分類付与済み特許文書の件数が 15 万件付近で正解率がほぼ頭打ちになっている．必要十分なデータ量は，分類の粒度やその規模に依存すると考えられるが，正解率のピークを 15 万件の時とした場合，1 件あたり平均 1.9 個のテーマが付与されているので，1 分類あたり平均約 1,000 件（150,000×1.9÷2,815＝約 1,000）の分類付与済み特許文書データがあれば必要十分であることが分かった．

表 2.3　自動生成されたテーマ分類知識に関する統計データ

項目	教師文書データ		
	公開特許公報 15 万件	分類付与マニュアル	両者を統合
ターム種類数	78,444 種類	51,036 種類	98,749 種類
分類知識レコード数	1,706,679 レコード	553,030 レコード	2,027,323 レコード
1 分類あたりのターム種類数	606 種類	196 種類	720 種類
最多タームを持つ分類のターム種類数	5,479 種類	1,968 種類	5,789 種類
最少タームを持つ分類のターム種類数	0 種類（26 分類）	1 種類	3 種類

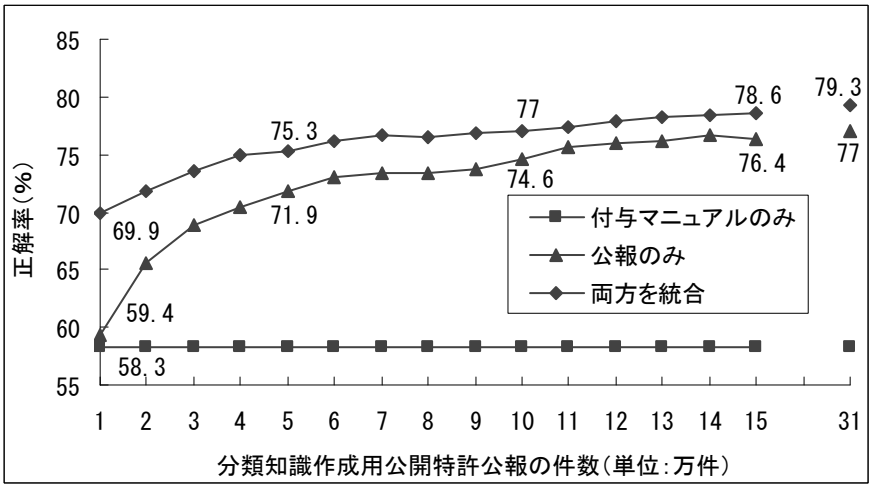


図 2.4　教師文書データの種類と規模の違いによる分類自動付与精度の比較（実験 1）

【実験 2】ターム重み付けに係るパラメータチューニングによる分類付与精度向上の検証

実験結果を**表 2.4** に示す．分類付与済み特許文書 15 万件と分類付与マニュアル文章を統合した分類知識を用いた場合（#5～#8）で比較すると，3 種類のチューニングにより分類付与精度を最大 5.3 ポイント（55.0%⇒60.3%）向上できた（表 2.4 の#5 と#8 のテーマ付与結果上位 1 位の正解率を比較）．特に正規化の効果が大きい（最大 2.3 ポイント（58.0%⇒60.3%）の精度向上）ことを確認した．上記チューニングパラメータの最適値は，教師文書データの量に依存するので，今回用いた最適値を必ずしも常に適用できるとは限らないが，最適化により精度を向上できることが分かった．しかし一方，審査室付与においては，正規化の効果がないことを確認した．この原因としては，審査室別の教師文書データの量に格差がそれほど見られず（最多と最少の格差が 4.0 倍にとどまった），正規化前後でタームの重み分布に違いが見られなかったことが挙げられる．なお，分類自動付与精度が最良である時（表 2.4 の#9）の各チューニングパラメータの最適値を**表 2.5** に示す．

表 2.4　3 種類の重みパラメータの最適化による分類自動付与精度の比較（実験 2）

#	分類知識生成方法		テーマ自動付与 正解率(%)				審査室自動付与 正解率(%)		
	教師文書データ	最適化するパラメータ	上位 1位	上位 3位	上位 5位	上位 10位	上位 1位	上位 2位	上位 3位
1	分類付与マニュアル	最適化しない	35.4	57.8	66.7	77.1	73.5	87.4	92.3
2	分類付与マニュアル	重みの上限を最適化	37.6	58.5	67.3	77.4	73.5	87.4	92.3
3	公開特許公報15万件	最適化しない	54.6	75.9	82.8	89.3	80.5	91.1	95.0
4	公開特許公報15万件	重みの上限を最適化	55.8	77.2	83.2	89.6	81.4	91.4	95.3
5	上記2種類を統合	最適化しない	55.0	76.8	83.8	90.0	80.7	91.6	95.3
6	上記2種類を統合	重みの上限を最適化	57.2	78.6	85.0	90.5	81.5	91.9	95.5
7	上記2種類を統合	重みの上限と統合方法を最適化	58.0	79.4	85.6	91.0	82.7	92.7	95.8
8	上記2種類を統合	重みの上限と統合方法を最適化し，更に正規化を施す	60.3	80.9	87.2	92.4	82.4	92.5	95.6
9	公開特許公報31万件と分類付与マニュアルを統合	重みの上限と統合方法を最適化し，更に正規化を施す	61.6	82.8	88.6	93.2	83.3	93.2	96.0

表 2.5　3 種類の重みパラメータの最適値（実験 2（表 2.4 の#9 の場合））

チューニングパラメータ	審査室付与	テーマ付与
分類別タームの重みに上限値を設ける（分類マニュアル）	1,000	30
分類別タームの重みに上限値を設ける（公開特許公報 31 万件）	30,000	1,500
タームの統合方法（定数 C の値の最適化）	20	15
重みの正規化（偏差値のしきい値の最適化）	正規化しない方が良い	200

表 2.6　大量の評価データによる分類自動付与精度評価（実験 3）

評価文書データ量	テーマ付与正解率(%)				審査室付与正解率(%)		
	上位 1 位	上位 3 位	上位 5 位	上位 10 位	上位 1 位	上位 2 位	上位 3 位
3,850 件	60.3	80.9	87.2	92.4	82.4	92.5	95.6
166,323 件	60.7	80.4	86.4	92.0	81.3	91.8	95.1

【実験 3】大量の評価データによる分類付与精度評価

実験結果を**表 2.6** に示す．テーマ自動付与精度は評価用特許文書件数が 3,850 件の時に比べて最大 0.8 ポイント（87.2%⇒86.4%），審査室自動付与では最大 1.1 ポイント（82.4%⇒81.3%）の正解率低下にとどまり，提案する分類自動付与方式は精度的に安定している

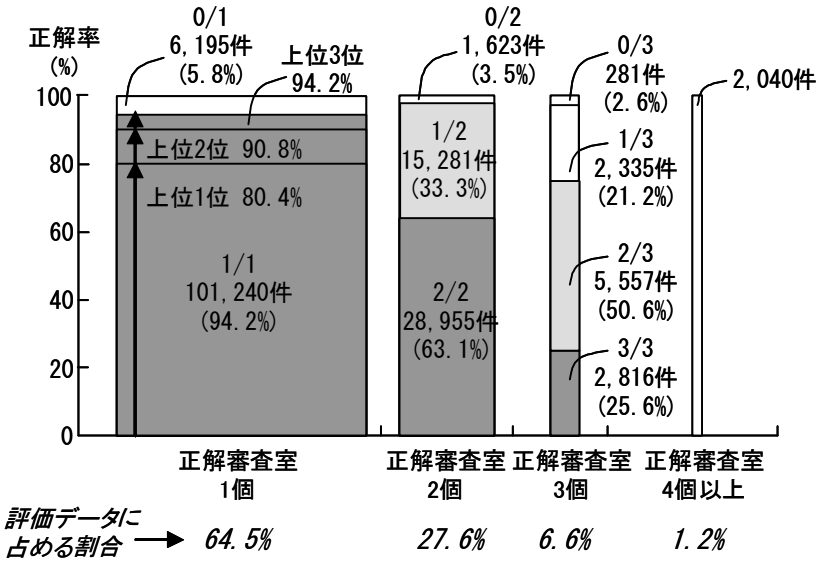


図 2.5　審査室分類付与の全自動化の実現可能性評価（実験 4）

ことが分かった.

【実験 4】全自動分類付与システムとしての実現可能性検証

実験結果を**図 2.5** に示す．棒グラフの面積が，該当する評価用特許文書件数に比例している（四つの棒の合計面積が評価用特許文書の全件数となる）ことに注意されたい．正解審査室を 1 種類だけ持つ評価用特許文書は全体の 64.5%であるが，そのうちの 94.2%については，分類自動付与結果として出力された上位 3 個の審査室の中に正解を含んでいた．また，評価用特許文書全体の 80.0%（濃い網掛けの部分の面積の総和，64.5%×94.2%+27.6%×63.1%+6.6%×25.6%+1.2%×0.0%）については，自動付与結果上位 3 個の審査室の中にすべての正解を含んでいた.

複数の分類が付与されるべき特許文書について，すべての分類を自動付与結果の上位に出力させることは，本研究のアプローチでは難しい．なぜなら，各分類を特徴付けるタームが，分類付与対象となる新規特許文書から必ずしも十分な数だけ得られる訳ではないからである．例えば，特許文書 A の正解分類が X と Y である場合，X，Y に関連するタームがそれぞれ十分な数だけ抽出できれば，X，Y の両方を自動付与することは本方式でも可能である．しかし，特許文書の記載内容の大部分が分類 X に関するものであり，分類 Y に関する内容が局所的または補足的である場合，その特許文書から抽出されるタームの多くは分類 X に関するものとなり，分類 Y に関するタームは相対的に少なくなる．その結果，分類

表 2.7　誤付与の原因分析

#	原因	該当件数
1	重要タームの照合漏れ	48
2	タームへの不適正な重み付け	42
3	複合語表現による照合漏れ	9
4	単語辞書の語彙不足	6
5	不要語除去による照合漏れ	4
6	構文・解析箇所による重み付けの失敗	2
7	表記の差異（発明者の記述ミス含む）	2
8	文章が短くタームが不足	1
9	その他	1

注：100 件対象，複数可

自動付与結果の上位には分類 X 及び分類 X に意味的に関連の深い分類が並んでしまい，分類 Y を上位に出力させることが困難となる．したがって，正解分類のすべてを自動付与結果の上位に出力させるためには，少数派の分類 Y を特徴付けるタームを特定する方式を検討するか，あるいは再現率が高い手法である KNN 法などと併用して付与すべき分類を特定する方式を検討する必要があると考える．

2.6.3　誤付与の原因分析

本節では，前節の評価実験において誤った分類が付与された評価用特許文書からランダムに抽出した 100 件について，その原因を手作業で分析した結果について述べる．**表 2.7** に分析結果を示すように，「重要タームの照合漏れ（分類知識中の正解分類を特徴付けるターム集合の中にその重要タームが含まれていない）」（48%）と，「タームへの不適正な重み付け」（42%）という 2 種類の原因が大半を占めていることが分かった．

重要タームの照合漏れを改善する方策としては，シソーラスや共起関係を用いたターム展開や部分文字列照合など，より柔軟なターム照合を行うことが考えられる．また，重み付けの不適正による誤付与を改善するためには，ターム別の重みのチューニングが必要である．また，ターム毎の分類別重み分布の偏りを解析して，そのタームが分類を特徴付けるのに有効なタームであるかを判別する方法も有効であると思われる．更に運用時におけるターム重み付けの学習方式は重要な研究課題である．

2.7 まとめ

本章では，出願特許文書に特許分類体系の上位分類である審査室及びテーマを自動付与する方式について述べた．特許文書は発明に係る「もの」またはものに対する「処理」を記述した文章であり，審査室及びテーマが「もの」または「処理」に基づいて網羅範囲を規定しているものが多いことを鑑み，本方式では，(1)特許文書の言語構造に着目し，特定の特許タグとして【発明の名称】,【請求項】,【技術分野】の 3 タグのみから発明の技術分野を特定するタームを抽出する方式，(2)特許文書におけるタームの出現位置及び出現共起に基づいてタームに重みを付与する方式，(3)大量の分類付与済み特許文書と，各分類の適用範囲を文章で規定した分類付与マニュアルからそれぞれ抽出したタームを統合し，各分類を特徴付ける重み付きターム集合からなる分類知識を全自動生成する方式，(4)タームの分類別出現傾向及び分類体系の階層性に着目した，特許文書と分類との間の類似度を算出する方式，を提案した．

本方式に基づき，最大約 31 万件の公開特許公報データ及び分類付与マニュアル文章から分類知識を自動生成し，新規特許文書にテーマ（2,815 分類）及び審査室（38 分類）を自動付与する精度評価実験を行った．その結果，テーマ及び審査室を 1 個ずつ自動付与した場合の正解率としてそれぞれ 61.6%，83.3%，3 個ずつ付与した場合の正解率としてそれぞれ 82.8%，96.0%を得た．また，分類付与済み特許文書データに加えて分類付与マニュアル文章を教師文書データとして利用することにより，正解率を 2.2 ポイントから 10.5 ポイントまでの範囲で向上でき，分類付与マニュアル文章を利用する有効性を確認した．更に分類知識保守の観点から，分類知識生成に必要十分な教師文書データ量を検証した結果，1 分類あたり約 1,000 件の分類付与済み特許文書があれば必要十分であることを確認した．

次に，審査室を対象として全自動分類付与システムとしての実現可能性を検証した結果，正解審査室を 1 個だけ持つ評価用特許文書（64.5%）のうちの 94.2%は，分類自動付与結果として出力された上位 3 個の審査室の中にその正解審査室を出力できること，評価用特許文書全体の 80.0%については，自動付与結果上位 3 個の審査室の中にすべての正解審査室を出力できることを確認した．

更に，誤った付与結果を出力した特許文書を解析してその原因を分析した結果，大半は重要タームの照合漏れまたはタームへの不適正な重み付けが原因であることが分かった．これらの原因を取り除き，精度を向上させる方式を検討することは今後の課題である．

次章及び第 4 章では，分類自動付与とともに特許審査作業の短縮化を実現するために必要な技術である類似特許文書検索の精度向上に関する研究について述べる．

第３章　文書内の言語構造を利用した類似特許文書検索

3.1　はじめに

出願特許の発明内容を無効化する過去の特許文書の検索精度向上は，特許庁における審査期間の短縮，企業における特許戦略のスピーディな立案/実践を実現する上で避けて通れない大きな技術課題である．

従来の特許文書検索システムでは，一つ以上のタームを検索条件（AND/OR/NOT や近傍検索などによる論理式）として指定し，その検索条件を満たす特許文書を類似度に相当するスコア順に出力するものが主流である．しかし，この類の特許文書検索では，検索条件として用いる適切なタームの選定及び論理式の組み立てが難しい上に，検索の意図に合致しないノイズ文書が検索結果に多く含まれることが多い．そのため，検索条件の作成には高度に知的なノウハウが不可欠となり，誰でも容易に特許文書を検索できないというのが現状である．実際に検索の専門家による検索のログを見ると，タームに加えて特許分類情報や出願人情報などを絡めた非常に複雑な検索条件による検索結果を適切に組み合わせることによって最終の検索結果を導出している事例が多い．

上記背景のもと，検索要求を文章として入力し，その内容に類似する文書を検索する自然言語文検索（類似文書検索）が特許文書検索においても注目されてきている．本方式は検索式を考える手間が大幅に軽減できるため，検索作業期間の短縮が期待できる．

一般に，特許の発明内容を最も端的に記載しているのは請求項である．したがって，自然言語文検索の入力として請求項文章を用いることにより，類似する特許文書を効率的に検索することができると考える．そこで本研究では，代表的な請求項である【請求項 1】タグの文章を入力として，そこに記載された発明内容に類似する過去の特許文書を高精度に検索する類似特許文書検索方式を提案する[54][55][56]．本方式は以下の 2 種類の処理方式を特徴としている．

(1) 類似特許文書検索を段階的に捉え，特許文書の言語構造を踏まえた上で，検索段階に応じてタームの抽出範囲及び検索対象を変え，各検索段階における検索結果から最終的な類似度を算出する二段階検索方式（3.2 節）

(2) 特許文書（特に請求項文章）の記述特性に着目して，タームを重み付けする方式（3.3 節）

以下，3.2 節では，上記二段階検索方式の特徴について詳細に述べる．3.3 節では，請求項の記述特性に着目した検索ターム重み付け方式として，タームの出現頻度を利用しないターム重み付け方式と，尺度を表すタームに着目したターム重み付け方式について述べる．3.4 節では，これらの方式に基づく類似特許文書検索方式の処理手順について述べる．3.5 節では，上記提案方式の有効性を検証する評価実験について述べ，実験結果について考察する．3.6 節では，類似特許文書検索研究の成果についてまとめる．

3.2　文書内の言語構造を利用した二段階検索方式

3.2.1　各段階における検索の特徴

本研究で提案する二段階検索方式の概要を**図 3.1** に示す．本方式は，1 回の検索指示に対して条件の異なる検索を 2 度実行し，これらの検索結果として得られる検索スコアをマージすることにより，最終の検索結果を得るものである．

第一段階は再現率（漏れ防止）重視の検索であり，検索結果上位 N 件の中により多くの類似特許文書を含めることを目的としたものである．ここでは，「広く浅い」類似文書検索方式を採用する．すなわち，入力請求項文章（以下，クエリと呼ぶ）の解析においては，クエリ全体からタームを抽出して重要度に相当する重みを付与し，検索においては，特許文書全文を検索範囲としている．228 件のクエリを用いた予備実験では，特許文書全文を検索範囲として検索を行った時に，検索結果上位 1,000 件に含まれる正解特許文書（検索されるべき類似特許文書）は 326 件であり，請求項のみを検索範囲とした場合（257 件）に比べて多くの正解特許文書を検索結果の中に集められることが確認できている．

第二段階は適合率（ノイズ防止）重視の検索であり，正解特許文書の順位を上げることを目的としたものである．第二段階では，第一段階で得られた検索結果の上位 N 件のみを検索対象とする．また，特許文書の構成及び請求項の文章構造を用いたクエリ解析及び検索方式を採用する．クエリ解析では，1.3.2 節で述べた請求項の構成要素のうち，特徴部分のみを解析対象とし，前提部分を無視することにより，発明の特徴（新規性または進歩性）を表すタームだけを検索に用いる．検索においても検索範囲を請求項のみに絞り，発明の特徴を表すタームをピンポイントに照合できるようにすることにより，正解特許文書の検索順位を向上させる．

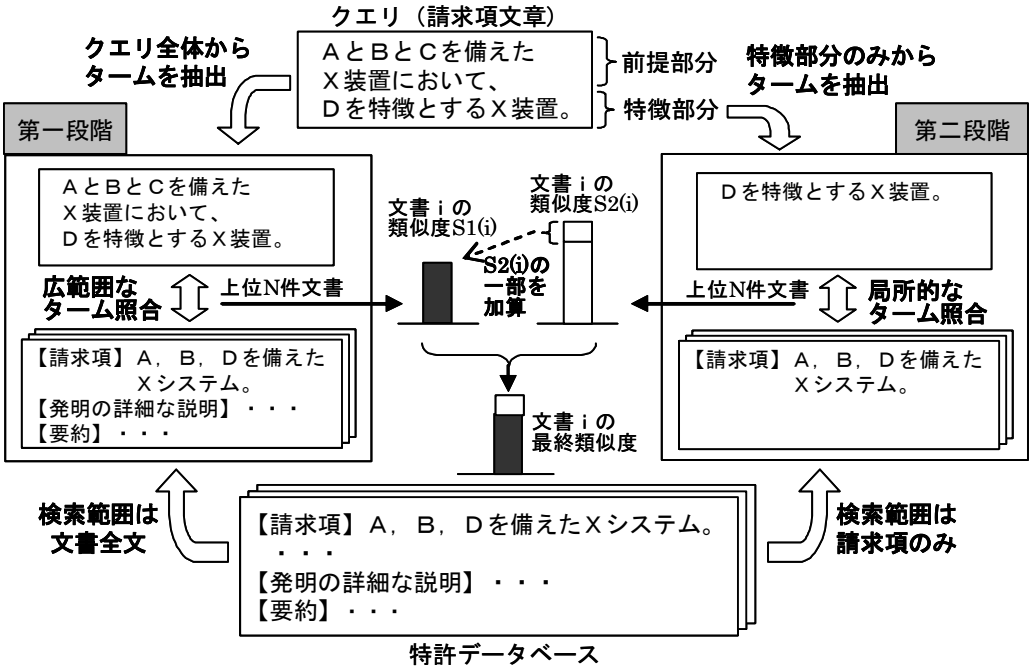


図 3.1　二段階検索方式の概要

3.2.2　類似度の算出方式

提案する二段階検索方式では，それぞれの検索段階で算出された類似度を最後にマージすることにより，最終の類似度を算出する．クエリと文書ｉの最終類似度 S(i)は以下の式により算出する．

$$\left.\begin{aligned} S(i) &= S1(i) + S2'(i) \times P \\ S2'(i) &= \frac{AVE(S1)}{AVE(S2)} \times S2(i) \\ AVE(X) &= \frac{\sum_{i=1}^{N} X(i)}{N} \end{aligned}\right\} \quad (3.1)$$

ここで，S1(i)は第一段階におけるクエリと検索結果文書ｉの類似度，S2(i)は第二段階におけるクエリと検索結果文書ｉの類似度，S2'(i)は第二段階におけるクエリと検索結果文書ｉの「正規化された」類似度，AVE(S1)は第一段階におけるクエリと検索結果文書ｉの類似度 S1(i)の平均値，AVE(S2)は第二段階におけるクエリと検索結果文書 ｉ の類似度 S2(i)の平均値，P は重みチューニング用パラメータ，N は検索結果として出力される文書件数である．予備実験の結果では，パラメータ P の最適値は 0.1 近辺であった．また，第

二段階の類似度 S2(i)を S2'(i)に正規化するのは，第一段階における類似度と第二段階における類似度は異なる方法で算出された値であり，類似度の価値が異なるためである．ここでは，類似度を統合する前に検索結果上位 N 件の文書の類似度の平均値を比較し，類似度の価値が同じになるように第二段階の類似度を補正する．

本算出方式の重要なポイントは，最終の類似度算出のベースとなる類似度は第一段階の類似度であり，第二段階の類似度は最終類似度に対して補完的に影響を及ぼす点であり，この点が従来の類似度算出方式と異なる点である．本方式の第二段階では，ターム抽出範囲及び検索範囲を狭めているため，ピンポイントでのターム照合により，検索順位が大きく改善される特許文書がある．しかしその一方で，重要なタームの欠落などの要因により，検索漏れが発生して検索順位が大きく悪化してしまう特許文書もある．第二段階での類似度を最終類似度としてそのまま用いると，ピンポイントでのターム照合による精度向上よりも，検索漏れによる精度悪化の方が強く働いてしまい，トータルでの精度は悪くなってしまうことが予備実験で確認されている．そこで，基本的には第一段階の検索結果を重視するが，特許文書の特性を活かした第二段階の検索結果を第一段階での検索結果に反映させることにより，請求項間の類似度の高い文書の順位を引き上げるという本方式を採用するに至っている．

3.3 請求項の記述特性を利用したターム重み付け方式

3.3.1 ターム出現頻度を用いないクエリターム重み付け方式

第一段階及び第二段階においてクエリから抽出されるタームへの重み付け方式として，TF-IDF 法をベースとした方式を用いる．TF-IDF 法では，あるクエリに出現するターム i の重みを次の式によって算出する．

$$W(i) = (1 + \log(TF(i)) \times \log\left(1 + \frac{N}{DF(i)}\right) \qquad (3.2)$$

ここで，TF(i)はクエリにおけるターム i の出現頻度，DF(i)は特許文書データベース中の特許文書の請求項（全文ではない）におけるターム i の出現文書数，N は特許文書データベース中の特許文書件数である．

本研究では，【請求項 1】の文章をクエリとして使う場合，ターム出現頻度（TF）はターム重み付けにとって不適切な手掛かりであると仮定する．これは以下の三つの根拠に基づいている．

(1) クエリとなる請求項文章が短い

請求項は階層構造をなしている．その中で最も発明内容を端的に述べているのは請求項の冒頭に記載される【請求項 1】である．本研究では，クエリとしてこの請求項 1 を用いるが，その文章長は短い．NTCIR-4 特許検索タスクのフォーマルランデータ（103 課題）の平均文章長は 340 バイト（170 字）であり，本研究における予備実験で用いた評価データ（228 課題）では 455 バイト（228 字）である．このように短い文章から得られるタームの TF がそのタームの重要度に比例しているとは考えにくい．

(2) 同じタームが繰り返し使われる

特許文書に記載された発明内容を人間が曖昧性なく正確に解釈できるように，請求項文章においては代名詞や「それ」，「これ」のような指示語を用いた曖昧な記述は使われない．その結果，同じタームが繰り返し使われることとなり，タームの TF がそのタームの本来の重要度以上の値となっていることがしばしばある．

(3) 発明対象を表すタームの TF が相対的に低くなる

図 1.3 で示した，請求項の末尾に現れる「発明対象」を表すタームは，発明の技術分野を絞り込む手掛かりとして重要なタームである．しかし，多くの請求項においては，発明対象を表すタームが請求項の末尾部分だけにしか出現しないことがしばしばあるため，ターム重み付けに TF を用いると，これらのタームの重要度が他のタームに比べて相対的に下がってしまう．

以上の根拠を鑑み，本方式では TF を用いないで出現文書数（DF）の情報だけからクエリ中のタームの重み付けをする方式を採用する．すなわち，式(3.2)において TF を 1 に固定した以下の算出式を用いてタームの重みを算出する．

$$W(i) = \log\left(1 + \frac{N}{DF(i)}\right) \qquad (3.3)$$

3.3.2 尺度表現に着目したクエリターム重み付け方式

第二段階におけるクエリターム重み付けでは，適合率を向上すべく，「尺度表現語」に着目する．尺度表現語とは，「速度」や「温度」など物性を表すタームであり，しばしば定量的属性値を伴うタームである．請求項において，数値条件によって発明の新規性または進歩性を主張する特許も少なくないので，尺度表現語は発明内容を特徴付けるタームであると考える．

日本語の尺度表現語は「圧縮率」，「類似度」など，単語末尾に特定の接尾語（「率」，「度」）

表 3.1　尺度表現語を特定する手掛かりとなる接尾語一覧

#	手掛かり接尾語	尺度表現語の例	#	手掛かり接尾語	尺度表現語の例
1	度	密度，温度	10	径	外径，長径
2	長	波長，全長	11	率	屈折率，力率
3	量	風量，水量	12	差	公差，誤差
4	数	件数，総数	13	温	室温，湯温
5	力	斥力，圧力	14	深	水深
6	高	残高，売上高	15	額	金額，総額
7	圧	電圧，浸透圧	16	比	S/N比，圧縮比
8	値	数値，閾値	17	幅	歩幅，線幅
9	速	風速，音速	18	点	融点，沸点

をしばしば伴う．そこで，これらの接尾語に着目して尺度表現語を漏れなく効率的に収集する．まず，単語辞書に登録された約 6 万語の基本単語の中から，**表 3.1** に示す 18 種類の接尾語のいずれかを末尾に持つ単語を抽出する．次に，抽出されなかった単語集合を人間が精査し，「pH」など接尾語を伴わない尺度表現語を抽出する．その結果，361 種類の尺度表現語を収集している．なお，表 3.1 に示されている接尾語そのものも尺度表現語として扱う．

「速度」や「温度」など，多くの尺度表現語は，さまざまな技術分野で使われるため，その尺度表現語の対象物を特定する（すなわち，「何の速度」，「何の温度」であるのかを特定する）ことが重要である．そこで，尺度表現語の対象物を表すタームを「尺度表現関連語」として定義する．尺度表現関連語はしばしば尺度表現語に隣接して出現する．そこで，助詞「の」を伴って尺度表現語を修飾しているタームまたは，尺度表現語とともに一つの複合語を形成しているタームを尺度表現関連語として抽出する．例えば，「用紙の搬送速度を制御する」という記述において，ターム「速度」は尺度表現語であり，ターム「用紙」及び「搬送」は尺度表現語「速度」を修飾しているので尺度表現関連語とみなす．第二段階では，上記の基準で抽出された尺度表現語及び尺度表現関連語に対して特定の重みを加算することにより，他のタームよりも重要度を高くする．

3.4　類似特許文書検索の処理フロー

前節までで述べた類似特許文書検索の処理フローを**図 3.2** に示す．第一段階では，以下の処理ステップに従って検索を実行する．これらの処理ステップは，従来のタームベース

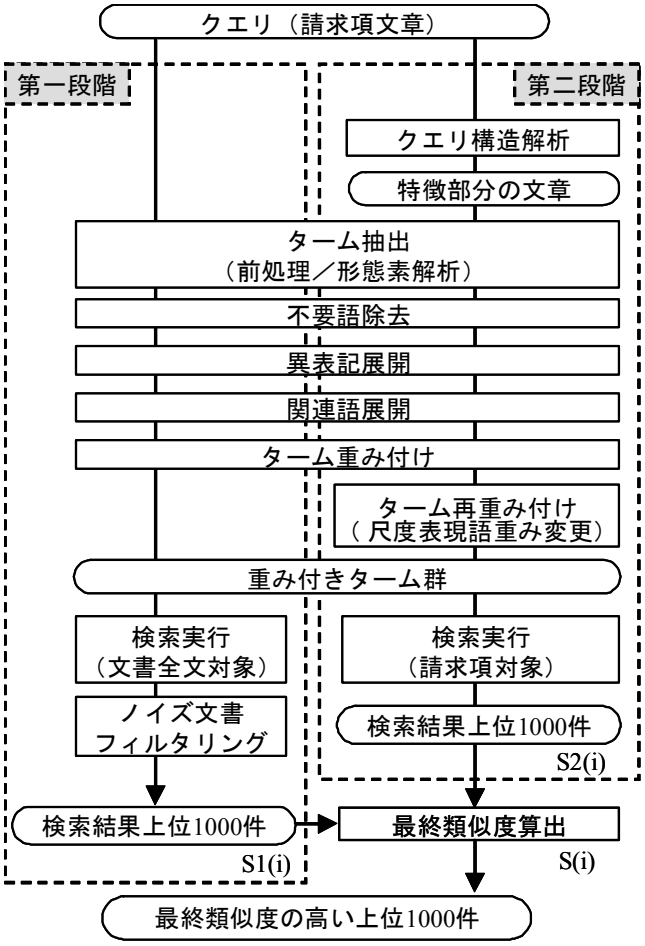


図 3.2　類似特許文書検索の処理フロー

の検索方式でも採用されている処理が多い.

(1) ターム抽出

単語辞書を参照してクエリを単語に分割し，品詞を割り当てる．単語分割には，フリーの形態素解析ツールである茶筌[7]を用いている[62]．なお，単語分割に先立ち，前処理として，日本語カタカナの直後に現れるハイフン“ー”はすべて日本語長音“ー”に変換しておく．次に，分割された単語の中から，名詞（サ変動詞語幹を含む），動詞（助動詞を含む），形容詞，アルファベット文字列を検索に使用するタームとして抽出する．

[7] 茶筌　http://chasen.naist.jp/hiki/ChaSen/

(2) 不要語除去

クエリから抽出されたタームの中から，重要でない（検索精度の向上に寄与しない）タームを除外する．本検索方式では，6 万語の基本語彙から予め人手で抽出した 2,910 語の不要語辞書を参照する．不要語には，「発明」や「装置」など，特許文書固有で頻繁に出現する単語と，「こと」や「する」など，特許文書以外の一般文書にも高頻度で出現する一般語を含む．

(3) 異表記展開

同じ意味であるが表記が若干異なるターム対（「インタフェース」と「インターフェイス」など）を同一のタームとして処理する．特許文書には多くのカタカナタームが存在するため，ここではカタカナタームのみを対象として異表記展開を行う．本処理では 1,365 種類の異表記展開ルールを展開時に使用する．一般に異表記展開は検索に必須の処理ではないが，3.5 節で述べる評価実験では，すべての実験パターンにおいて本処理を適用している．

(4) 関連語展開（オプション）

大量の特許文書を自然言語解析して二つの任意のターム間の意味的関連性を算出し，関連性の強いターム対を特定して関連語辞書に予め登録蓄積しておく．そして，この関連語辞書を参照することによって，クエリから抽出されたタームを拡張する．本処理によって拡張されたタームにより，ターム照合がより漏れなくなされるため，再現率が改善されると考える．関連語辞書は，タームの共起性及び【符号の説明】タグにおける括弧表現を手掛かりとして全自動で生成される[55][56]．

(5) ターム重み付け

式(3.3)に基づき，検索対象となる特許文書から算出される IDF 値をクエリタームの重みとして割り当てる．ここで，クエリが請求項文章であることから，DF 算出の解析範囲は特許文書全文ではなく，請求項文章のみとしている．

(6) 検索実行

クエリに類似する特許文書を特許文書データベースから検索する．そして，類似度の高い上位 1,000 件の特許文書を出力する．3.2.1 節で述べたように，第一段階での検索範囲は特許文書全文である．なお，検索エンジンとして，汎用連想検索エンジン GETA[8]を用いている[105]．

---

[8] GETA: http://geta.ex.nii.ac.jp/ 「汎用連想検索エンジン (GETA)」は，情報処理振興事業協会 (IPA) が実施した「独創的情報技術育成事業」の研究成果である．

(7) ノイズ文書フィルタリング（オプション）

検索結果に含まれるノイズ文書を極力削除すべく，国際特許分類（IPC）の利用，部分ターム集合の利用，パッセージの利用という 3 種類の文書フィルタリング方式を採用する[55]．一般に，文書フィルタリングは再現率重視というよりは適合率重視の処理である．しかし，本研究で提案する二段階検索方式においては，第一段階の検索結果上位 1,000 件に多くの類似特許文書を含めることが重要である．ノイズ文書フィルタリングは，検索結果出力順位のしきい値以下にある正解特許文書を拾い上げるのに有効であるが，その一方で，正解特許文書の検索結果が 1 位であってもフィルタリング条件を満たさないと除外されてしまい，検索精度を大きく低下させてしまうというトレードオフの関係を持つ．予備実験結果において本フィルタリング処理を適用することにより，検索結果上位 1,000 件に含まれる正解特許文書が 326 件から 334 件に増加したことから，本処理を第一段階で使用している．

第二段階では以下の処理ステップに従って検索を実行する．第二段階における処理の多くは第一段階のそれと共通である（前処理及び形態素解析，不要語除去，異表記・関連語展開など）．第一段階と異なる処理は以下の三つである．

(1) クエリ構造解析

クエリとなる請求項文章を解析して，請求項の前提部分と特徴部分の境界を特定する．日本語請求項では，境界点において，後置詞表現「～において,」または「～であって,」が使われることがほとんどであるので，これらの表層表現を手掛かりに境界を決定する．なお，上記手掛かり表現が一つのクエリの中に複数出現する場合，最も後ろに出現する箇所を境界とする．第二段階では請求項の特徴部分のみをターム抽出範囲とするので，本処理の出力は特徴部分の文章である．

(2) ターム再重み付け

クエリに出現する尺度表現語及び尺度表現関連語を特定し，これらに重みを加算する．ここでは，該当するタームの TF をそれぞれ 2 ずつ増加させ（したがって TF は 1+2=3 となる），式(3.2)によって重みの値を再計算する．

(3) 検索実行

第一段階と同一の検索アルゴリズムが適用される．第一段階の検索との違いは，クエリタームが異なることと，検索範囲が請求項文章のみに限定されていることである．

最後に第一段階の検索結果と第二段階の検索結果から，各文書の最終類似度（S(i)）を式(3.1)を用いて算出する．

3.5 検索精度評価実験

3.5.1 実験方法

(1) 実験データ

本実験では，約 170 万件の公開特許公報データ（1993 年～1997 年公開）を特許文書データベースとする．クエリとしては，NTCIR-4 特許検索タスクのフォーマルランの課題データを使用する．本データは，メインデータ（以下，MAIN）と追加データ（以下，ADD）という 2 種類のデータセットからなる．MAIN は 34 課題からなり，正解特許文書は 344 件（課題 1 件あたり 10.1 件）である．MAIN の正解特許文書は，その技術分野の専門家が人手で収集したものである．一方，ADD は実質 69 課題からなり，正解特許文書は 115 件（課題 1 件あたり 1.67 件）である．ADD の正解特許文書は，特許庁における実際の特許審査において審査官が発明を無効化するために引用した特許文書である．

正解特許文書は二つの集合 a，b に分けられる．集合 a はその正解特許文書単独でその課題の発明を無効化できる正解特許文書集合であり，集合 b は他の正解特許文書と組み合わせることによってはじめて課題の発明を無効化できる正解特許文書集合である．本実験では，集合 a と集合 a+b という二つの正解特許文書集合を用いて検索精度を評価する．そして，以下では集合 a に対応する評価データセットを MAIN-a，ADD-a と呼び，集合 a+b に対応する評価データセットを MAIN-ab，ADD-ab と呼ぶこととする．なお，本研究の予備実験で用いた 228 課題（正解特許文書 460 件）からなる評価データセット（以下，PRE と呼ぶ）も本実験で使用する．この評価データセットは，1999 年 1 月に公開された特許の一部で構成されており，正解特許文書は，特許庁がインターネット上で運営する特許電子図書館[9]（IPDL: Industrial Property Digital Library）から収集したものであり，上記集合 a+b（PRE-ab）に相当する．

(2) 検索結果の出力

特許文書 170 万件から，各課題に対する検索結果上位 1,000 件を全自動で出力する．また，出願日情報が考慮され，課題特許が出願された出願日よりも前に公開された特許文書のみが検索対象となる．

検索精度の評価尺度として，NTCIR-4 で採用された「平均精度の課題別平均（Mean Average Precision，以下 MAP と呼ぶ)」を採用する．すなわち，1 件の課題に対する平均

---

9 特許電子図書館　http://www.ipdl.ncipi.go.jp/homepg.ipdl
現在は特許庁からの委託を受けた独立行政法人工業所有権情報・研修館によって運営されている．

精度を以下の式で算出し，課題毎の平均精度の平均値を求める．

$$\text{平均精度}=\frac{1}{\sum_{i=1}^{N} X_i}\sum_{i=1}^{N}\left[\frac{X_i}{i}\left(1+\sum_{k=1}^{i-1} X_k\right)\right] \qquad (3.4)$$

ここで，N は出力される公開特許公報件数（本実験では N=1,000），Xi は i 位に出力された公報が正解特許文書であるかを示す値（正解特許文書なら 1，それ以外は 0）である．また MAP とは別の評価尺度として，「ある検索精度向上方式を適用することによって検索順位がベースライン（その方式を適用しない場合）よりも向上した正解特許文書件数の割合」を適宜導入する．

(3) 実験内容

本実験では，3.2 節で提案した「二段階検索方式」及び，3.3 節で提案した「ターム出現頻度を用いないクエリターム重み付け方式」「尺度表現に着目したクエリターム重み付け方式」の 3 種類の方式に焦点を絞ってその有効性を検証する．なお，すべての実験において不要語除去処理及び異表記展開処理が適用されている．

3.5.2 実験結果と考察

(1) ターム出現頻度を用いないクエリターム重み付け方式の有効性検証

本方式の評価結果を**表 3.2** に示す．表 3.2 では，評価データセット毎に，本方式を適用しない場合と適用した場合の MAP 値を表の最下段に示している．また，各クエリにおいて，

表 3.2　ターム出現頻度を利用しないクエリターム重み付け方式の有効性評価

TFの最大値	PRE					MAIN							ADD						
	クエリの数			MAP(PRE-ab)		クエリの数			MAP(MAIN-a)		MAP(MAIN-ab)		クエリの数			MAP(MAIN-a)		MAP(MAIN-ab)	
	件数	割合(%)	割合累計(%)	TFN	TF1	件数	割合(%)	割合累計(%)	TFN	TF1	TFN	TF1	件数	割合(%)	割合累計(%)	TFN	TF1	TFN	TF1
1	2	0.9	0.9	0.0009	0.0009	1	2.9	2.9	0.0476	0.0476	0.3342	0.3342	4	6.0	6.0	0.2571	0.2571	0.2571	0.2571
2	20	8.8	9.6	0.1614	0.1639	7	20.6	23.5	0.1380	0.1421	0.2064	0.2223	10	14.9	20.9	0.2088	0.1417	0.2093	0.1465
3	25	11.0	20.6	0.1496	0.1987	5	14.7	38.2	0.3422	0.3822	0.1784	0.2080	13	19.4	40.3	0.1328	0.1311	0.1251	0.1219
4	31	13.6	34.2	0.2139	0.2252	7	20.6	58.8	0.3921	0.3854	0.3199	0.3343	7	10.4	50.7	0.1375	0.1535	0.1387	0.1573
5	23	10.1	44.3	0.1446	0.1286	3	8.8	67.6	0.1987	0.3484	0.1485	0.2473	8	11.9	62.7	0.0500	0.0295	0.0517	0.0352
6	28	12.3	56.6	0.1155	0.1401	2	5.9	73.5	0.1250	0.1346	0.0954	0.1028	7	10.4	73.1	0.0429	0.1326	0.0429	0.1326
7	19	8.3	64.9	0.2177	0.2770	1	2.9	76.5	0.1222	0.2500	0.0611	0.1285	6	9.0	82.1	0.0642	0.0977	0.0642	0.0977
8	21	9.2	74.1	0.0486	0.0861	1	2.9	79.4	–	–	0.0134	0.0182	2	3.0	85.1	0.0092	0.0089	0.0093	0.0089
9	15	6.6	80.7	0.0798	0.0719	0	0.0	79.4	–	–	–	–	1	1.5	86.6	0.0000	0.0013	0.0000	0.0013
10-	44	19.3	100	0.1282	0.1450	7	20.6	100	0.2043	0.2218	0.2297	0.2441	9	13.4	100	0.0650	0.0392	0.0650	0.0392
合計	228	100	100	0.1411	0.1607	34	100	100	0.2416	0.2696	0.2142	0.2393	67	100	100	0.1100	0.1088	0.1107	0.1097

注：網掛けはTF1とTFNのMAPの良い方に施されている

表 3.3　検索順位の改善/悪化の観点から見た各方式の有効性評価

#	処理方式	ベースラインの検索順位	処理方式を適用した時に検索結果上位1,000件に含まれる正解特許文書の順位変動（件）														
			PRE-ab			MAIN						ADD					
						MAIN-a			MAIN-ab			ADD-a			ADD-ab		
			順位改善	順位悪化	変動なし	順位改善	順位悪化	変動なし	順位改善	順位悪化	変動なし	順位改善	順位悪化	変動なし	順位改善	順位悪化	変動なし
1	ターム出現頻度を用いないターム重み付け（表3.2）	1- 10	26	31	43	13	11	17	28	19	29	5	5	6	5	5	7
		11- 100	44	56	5	29	22	3	60	48	4	9	16	1	11	18	1
		101-1000	57	74	0	16	8	0	46	32	1	13	14	7	15	20	7
		合計	127	135	48	58	38	20	134	99	34	27	35	14	31	43	15
2	二段階検索（表3.4 #3-4）	1- 10	27	13	59	11	9	26	21	16	46	3	8	8	3	8	9
		11- 100	64	30	9	26	27	1	61	46	6	11	6	1	12	8	2
		101-1000	64	59	0	11	9	0	46	31	0	26	13	0	32	15	0
		合計	155	102	68	48	45	27	128	93	52	40	27	9	47	31	11
3	尺度表現に着目したターム重み付け（表3.6 #2-5）	1- 10	6	3	97	4	1	41	4	5	80	1	3	13	1	3	14
		11- 100	22	12	66	10	10	29	34	17	50	10	0	14	10	0	18
		101-1000	39	15	65	15	0	10	36	10	37	12	3	20	17	4	22
		合計	67	30	228	29	11	80	74	32	167	23	6	47	28	7	54

注：網掛けされた値は順位が改善／悪化した正解特許件数の多い方を示す

各タームが持つ TF の最大値によって，クエリをグルーピングした時のグループ別の MAP を示している．

表 3.2 最下段の MAP 値が示すように，評価データセット MAIN と PRE において本方式は非常に有効である．すなわち，本方式を適用する（以下，TF1 と呼ぶ）ことにより，適用しない（以下，TFN と呼ぶ）場合に比べて，MAP は 11.6%（MAIN-a において 0.2416→0.2696）から 13.9%（PRE において 0.1411→0.1607）改善される．一方，評価データセット ADD では MAP はほとんど変わらない．

一方，各タームが持つ TF の最大値によって，クエリをグルーピングした時のグループ別の MAP を比較すると，MAIN 及び PRE においては，ほとんどのグループで TF1 の方が MAP が高くなっている．特に，TF の最大値が 5 より大きいグループにおいて改善の度合が大きい．これに対して ADD では，およそ半分のグループで TF1 の方が MAP が低くなっている．特に，TF の最大値が 2 から 3 において MAP の低下が顕著である．その原因の一つとして，ADD において TF の最大値が 3 以下のグループに属するクエリ件数の割合（40.3%）が PRE（20.6%）及び MAIN（38.2%）に比べて高くなっており，TF を 1 に固定する効果が十分に出ていないクエリが多いことが挙げられる．

表 3.3 は検索順位の改善/悪化の観点から，各方式の有効性を評価した結果である．このうち#1 は，本方式の適用により順位を上げた正解特許文書件数の割合を示している．MAIN-a 及び MAIN-ab では割合がそれぞれ 60.4%（58/(58+38)），57.5%（134/(134+99)）と高い一方， ADD-a，ADD-ab 及び PRE では 50%以下となっている（それぞれ 43.5%，41.9%，48.5%）．

表 3.2 及び表 3.3 の結果から，本方式は検索の適合率を向上させる方式としては有効であるが，検索順位を全体的に押し上げる方式としては十分に有効ではないことを確認した．

(2) 二段階検索方式の有効性検証

本方式の評価では，以下の 4 種類の選択肢の組み合わせを考慮しなければならない．

(a) クエリタームの抽出範囲　：　クエリ全体が良いか，特徴部分のみで良いか
(b) 検索範囲　：　特許文書全文が良いか，請求項文章のみで良いか
(c) 類似度スコア算出の対象文書：　全文書か，第一段階の検索結果上位文書のみか
(d) 類似度スコア算出　：　第一段階の類似度か，第二段階の類似度か，両者を統合するか

表 3.4 は，上記選択肢のどの組み合わせが有効かに関する精度評価実験結果を示している．まず，表 3.4 の#1 が示すように，クエリターム抽出範囲を特徴部分のみに限定しても（#1-3, #1-4），検索範囲を請求項のみに限定しても（#1-2, #1-4），ベースライン（#1-1）に比べて MAP は向上しない．特に，検索範囲を請求項のみに限定すると，MAP は極端に悪くなる．次に表 3.4 の#2 が示すように，類似度算出の対象文書を第一段階の検索結果上位 1,000 件に限定すると，MAP の値は若干改善されるが，それでもベースライン（#1-1）の値には届かない．

これらに対して，表 3.4 の#3 が示すように，本研究で提案する類似度算出方式(式(3.1))を適用すると，MAP はベースラインよりも良くなり，最大で 6.0%（#3-4 の PRE で 0.1607

表 3.4　二段階検索方式の有効性評価

#		検索方式（段階数）	解析方法				MAP				
			第一段階		第二段階			MAIN		ADD	
			タ－ム抽出範囲	検索範囲	タ－ム抽出範囲	検索範囲	PRE-ab	MAIN-a	MAIN-ab	ADD-a	ADD-ab
1	1-1	一段階検索（第一段階のスコア）	請求項	全文	－	－	0.1607	0.2696	0.2393	0.1088	0.1097
	1-2			請求項のみ	－	－	0.1195	0.1062	0.1012	0.0419	0.0428
	1-3		特徴部分のみ	全文	－	－	0.1569	0.2349	0.2314	0.1003	0.1036
	1-4			請求項のみ	－	－	0.1195	0.0953	0.1046	0.0347	0.0361
2	1-1	二段階検索（第二段階のスコア）	請求項	全文	請求項	全文	0.1607	0.2696	0.2393	0.1088	0.1097
	2-2					請求項のみ	0.1205	0.1663	0.1680	0.0835	0.0833
	2-3				特徴部分のみ	全文	0.1558	0.2351	0.2319	0.1005	0.1037
	2-4					請求項のみ	0.1211	0.1552	0.1688	0.0767	0.0772
3	1-1	二段階検索（式3.1でスコア算出）	請求項	全文	請求項	全文	0.1607	0.2696	0.2393	0.1088	0.1097
	3-2					請求項のみ	0.1654	0.2626	0.2419	0.1114	0.1119
	3-3				特徴部分のみ	全文	0.1640	0.2698	0.2402	0.1090	0.1098
	3-4					請求項のみ	0.1703	0.2660	0.2433	0.1124	0.1130

注：網掛けの数値は各評価セットでの最良値を示している

表 3.5　二段階検索方式におけるパラメータ P の最適化

Pの値	MAP				
	PRE-ab	MAIN		ADD	
		MAIN-a	MAIN-ab	ADD-a	ADD-ab
0.000	0.1607	0.2696	0.2393	0.1088	0.1097
0.025	0.1667	0.2706	0.2454	0.1178	0.1184
0.050	0.1695	0.2704	0.2452	0.1141	0.1147
0.075	0.1714	0.2663	0.2429	0.1126	0.1132
0.100	0.1703	0.2660	0.2433	0.1124	0.1130
0.125	0.1702	0.2646	0.2426	0.1059	0.1078
0.150	0.1693	0.2613	0.2420	0.1038	0.1058
0.200	0.1587	0.2527	0.2383	0.1023	0.1046
0.300	0.1548	0.2403	0.2314	0.0975	0.1000
0.400	0.1534	0.2386	0.2294	0.0941	0.0969
0.600	0.1479	0.2339	0.2250	0.0983	0.1010
0.800	0.1447	0.2318	0.2211	0.0957	0.0986
1.000	0.1397	0.2254	0.2165	0.0906	0.0926

注:網掛けの数値は各評価セットの最良値である

→0.1703）向上する．注目すべきは，二段階検索方式を適用する場合，ターム抽出範囲を特徴部分に絞っても，また検索範囲を請求項のみに限定しても，どちらも MAP が向上する点である．このことは，MAIN-a 以外のすべての評価データセットにあてはまる．

表 3.3 の#2 は本方式の適用により順位を上げた正解特許文書件数の割合を示している．順位が変動した正解特許文書のうちの 57～60%が順位を上げており，本尺度からも二段階検索方式が有効であることが分かる．

表 3.5 は式(3.1)におけるパラメータ P の値を変えた時の MAP の値を比較したものである．本評価実験では，P＝0.1 としているが，P の最適値は，PRE では 0.075，MAIN 及び ADD では 0.025 あたりであることが分かる．クエリターム抽出範囲を特徴部分のみに限定しても検索精度は悪化しないが，検索範囲を請求項文章のみに限定すると検索精度は大きく悪化する．この検索精度が第一段階での検索精度に匹敵するほど改善されれば，上記 P の値は大きくなる，すなわち第二段階での検索結果が最終検索結果に与えるプラスの影響が大きくなると考えられる．

(3) 尺度表現に着目したクエリターム重み付け方式の有効性検証

本方式の評価結果を**表 3.6** に示す.表内の MAP 値で,括弧内の数値は比較対象となる MAP 値を表し，括弧外の数値はその左側に記載されている条件で本方式を適用した場合の MAP 値を表している．括弧内外の MAP の値を比較すると，本方式は上記二つの方式に比べて，MAP の改善にあまり貢献していないことが分かる．この理由としてはまず，尺度表現語は

表 3.6　尺度表現に着目したクエリターム重み付け方式の有効性評価

#	検索方式（段階数）	第一段階			第二段階				MAP				
		ターム抽出範囲	検索範囲	尺度表現	ターム抽出範囲	検索範囲	尺度表現	式(3.1)のスコア算出方式	PRE-ab	MAIN		ADD	
										MAIN-a	MAIN-ab	ADD-a	ADD-ab
1	一段階検索	請求項	全文	適用	-	-	-	-	0.1600 (0.1607)	0.2705 (0.2696)	0.2403 (0.2393)	0.1062 (0.1088)	0.1069 (0.1097)
2-1	二段階検索	請求項	全文	-	請求項	全文	適用	-	0.1600 (0.1607)	0.2708 (0.2696)	0.2407 (0.2393)	0.1062 (0.1088)	0.1070 (0.1097)
2-2				-	請求項	請求項のみ	適用	-	0.1118 (0.1205)	0.1667 (0.1663)	0.1620 (0.1680)	0.0763 (0.0835)	0.0763 (0.0833)
2-3				-	特徴部分のみ	全文	適用	-	0.1525 (0.1558)	0.2373 (0.2351)	0.2330 (0.2319)	0.0980 (0.1005)	0.1012 (0.1037)
2-4				-	特徴部分のみ	請求項のみ	適用	-	0.1123 (0.1211)	0.1580 (0.1552)	0.1653 (0.1688)	0.0701 (0.0767)	0.0707 (0.0772)
2-5				-	特徴部分のみ	請求項のみ	適用	適用 (P=0.1)	0.1697 (0.1703)	0.2683 (0.2660)	0.2443 (0.2433)	0.1095 (0.1124)	0.1102 (0.1130)

注１：括弧内の数値はベースライン（比較対象）となる方式のMAP値
注２：網掛けの数値は、ベースラインよりMAP値が高いもの

表 3.7　評価実験結果の総括

#	評価尺度	提案した検索方式	方式の有効性				
			PRE-ab	MAIN		ADD	
				MAIN-a	MAIN-ab	ADD-a	ADD-ab
1	MAP	ターム出現頻度を用いないターム重み付け（表3.2）	◎(13.9%)	◎(10.4%)	◎(11.7%)	△(-1.1%)	△(-0.9%)
		二段階検索（表3.4）	○(6.7%)	△(0.4%)	△(2.5%)	◎(8.3%)	◎(7.9%)
		尺度表現に着目したターム重み付け（表3.6）	△(-0.4%)	△(0.9%)	△(0.4%)	△(-2.6%)	△(-2.5%)
2	検索順位の改善／悪化	ターム出現頻度を用いないターム重み付け（表3.3 #1）	△(48.5%)	○(60.4%)	○(57.5%)	×(43.5%)	×(41.9%)
		二段階検索（表3.3 #2）	○(60.3%)	△(51.6%)	○(57.9%)	○(59.7%)	○(60.3%)
		尺度表現に着目したターム重み付け（表3.3 #3）	◎(69.1%)	◎(72.5%)	◎(69.8%)	◎(79.3%)	◎(80.0%)

注1：#1では、◎はベースラインに比べて7%以上良い、○は3%～7%良い、△は-3%～3%良い、×は3%以上悪いことを示す
注2：#2では、◎は順位変動件数に占める順位改善件数の割合が65%以上、○は55%～65%、△は45%～55%、×は45%以下であることを示す

全ての課題に出現する訳ではなく，MAIN，ADD，PRE でそれぞれ，50.0%（17 件/34 件），55.1%（38 件/69 件），49.1%（112 件/228 件）のクエリにしか出現していないことが挙げられる．第二の理由としては，本方式では課題から抽出されるタームのうち，数種類のタームの重みが多少変わるだけであるため，類似度スコアに与える影響の度合が比較的小さいことが挙げられる．第三の理由としては，尺度表現語として用いるタームの表記が人によって異なっていることが挙げられる（例えば，「速度」と「スピード」）．尺度表現語の語彙数は少ないので，その同義語まで含めて尺度表現語リストを整備することは比較的容易であると考える．

一方，表 3.3 の#3 が示すように，本方式の適用により順位を上げた正解特許文書件数の割合は，69%～80%と格段に高くなっており，本方式が検索順位を全体的に押し上げるのに非常に有効であることが分かった．特に，ベースラインの正解特許文書の検索順位が 11

表 3.8　検索結果出力例

クエリ請求項文章(NTCIR-4 #40(公開番号1999-084093))				
軽元素層と重元素層を交互に積層してなるX線用多層膜光学素子において、前記軽元素層にフラーレンを用いたことを特徴とするX線用多層膜光学素子。				
類似特許文書検索結果				
順位	類似度	公開番号	発明の名称及び請求項1	正解
1	0.2513	1994-230194	【発明の名称】X線反射鏡 【請求項1】複数の層対が積層され、ブラッグ回折効果を有する多層膜構造のX線反射鏡において、 上記多層膜の軽元素層中にフラーレンを含むことを特徴とするX線反射鏡。	○
2	0.2403	1994-242297	【発明の名称】分光素子 【請求項1】金属内包フラーレン結晶により構成されたことを特徴とする分光素子。	○
3	0.2361	1993-066296	【発明の名称】軟X線多層膜分光素子 【請求項1】重元素層と軽元素層とが交互に積層されたブラッグ回析効果を有する多層膜を備えた 軟X線多層膜分光素子において、前記重元素層と軽元素層とが接する積層面に対して平行でない 多層膜の表面部を有し、この表面部によって最外面が形成されていることを特徴とする軟X線多層膜分光素子。	
4	0.2299	1996-327795	【発明の名称】X線用シュヴァルツシルト型光学系およびこれを用いる元素マッピング方法 【請求項1】X線の反射率を向上するために、重元素と軽元素とを積層してなる多層膜をそれぞれ 設けた多層膜凹面鏡および多層膜凸面鏡を有するX線用シュヴァルツシルト型光学系において、 前記軽元素の吸収端の波長をλ、前記多層膜凹面鏡および多層膜凸面鏡へのX線の入射角の 最大値をθMAX としたとき、前記多層膜の1周期の厚さdが、d＜λ／(2 cosθMAX)を満たし、 かつ、前記軽元素の吸収端前後の波長領域で、それぞれ少なくとも一つの透過率のピークを有する よう構成したことを特徴とするX線用シュヴァルツシルト型光学系。	
5	0.2248	1997-236696	【発明の名称】多層膜分光素子 【請求項1】反射層と、この反射層を形成する元素よりも原子番号の小さい元素を含むスペーサ層とを 交互に複数組積層して構成され、入射X線を回折して回折X線を発生させる多層膜分光素子において、 X線が入射される表面に、前記反射層を形成する元素よりも原子番号の小さい元素を含む干渉層が形成され、 この干渉層の厚みは、前記入射X線に含まれた回折対象のX線に波長が近いバックグラウンドX線について、 前記干渉層の表面での反射X線と、最上層の反射層からの反射X線とが打ち消し合う位相になるように 設定されていることを特徴とする多層膜分光素子。	
6	0.2244	1994-308308	【発明の名称】多層膜分光素子 【請求項1】ブラッグ回折効果を有する多層膜分光素子において、多層膜と基板との界面に緩衝層を 形成させたことを特徴とする多層膜分光素子。	○
7	0.2184	1995-005296	【発明の名称】軟X線用多層膜 屈折率の異なる2種の薄膜がそれぞれ所定の膜厚で交互に積層してなる軟X線用多層膜において、 屈折率の低い方の薄膜が、主としてコバルト(Co)とクロム(Cr)の合金より成り、しかも該合金の組成式が CoxCr1-x(式中、xは0．3ないし0．8の数字である。)であることを特徴とする軟X線用多層膜。	
8	0.2179	1994-230193	【発明の名称】X線分光反射鏡 【請求項1】複数の層対が積層された多層膜構造をもち、ブラッグ回折効果を有するX線分光反射鏡において、 上記積層中の少なくとも1層に弗素を含むことを特徴とするX線分光反射鏡。	
9	0.2166	1994-273596	【発明の名称】X線光学素子 【請求項1】重元素層と軽元素層とを交互に積層して構成されるブラッグ回折効果を有する多層膜分光素子 において、上記重元素層はCoもしくはCoを主成分とする化合物よりなることを特徴とするX線光学素子。	○
10	0.2155	1993-203798	【発明の名称】多層膜分光反射鏡 【請求項1】ブラッグ回折効果を有する多層膜分光素子の重元素層と軽元素層との間にSiとCからなる 化合物中間層を使用したことを特徴とする多層膜分光反射鏡。	○
・・				
37	0.1907	1996-005795	【発明の名称】軟X線多層膜反射鏡 【請求項1】軟X線に対する屈折率が高い第1の物質と、それより屈折率の低い第2の物質とが交互に 積層される多層膜構造を有し、前記第1の物質より成る高屈折率層の厚さと、前記第2の物資より成る 低屈折率層の厚さとが、それら両層の複数の境界でそれぞれ反射される軟X線がお互いに強めあうように 設定されている軟X線多層膜反射鏡において、前記第1の物質が炭素のフラーレン分子構造を有する ことを特徴とする軟X線多層膜反射鏡。	○
・・				
59	0.1801	1997-230098	【発明の名称】多層膜X線反射鏡 【請求項1】複数の物質層を周期的に積層した多層膜X線反射鏡において、上記物質層の各層間に 中間層を形成し、上記中間層として、少なくとも一つの上記物質層よりも融点の高い物質を使用する ことを特徴とする多層膜X線反射鏡。	○
・・				
71	0.1766	1996-122496	【発明の名称】多層膜反射鏡 【請求項1】軟X線領域での屈折率と真空の屈折率との差が小さい物質の第1層と大きい物質の第2層とを 基板上に交互に積層してなる多層膜反射鏡において、前記屈折率の差が小さい物質として、ホウ素、炭素、 炭化ホウ素、または窒化ホウ素を用い、前記屈折率の差が大きい物質として、ルテニウム、ロジウム、 パラジウム、または銀を用いたことを特徴とする多層膜反射鏡。	○
・・				
110	0.1667	1994-184738	【発明の名称】炭素薄膜の形成方法とその改質方法およびその改質方法を用いた電子デバイスおよび X線多層膜ミラーとその製造方法 【請求項1】基板上に、多面体、円筒またはらせん形のいずれか1種の形状を有する複数の炭素原子 からなる分子を薄膜状に形成する際に、前記分子または前記分子以外の分子または原子を一つ以上 イオン化し加速して蒸着することを特徴とする炭素薄膜の形成方法。	○
・・				

位から 1,000 位の課題に本方式を適用した場合の改善の度合が強い．逆に，1 位から 10 位の課題では順位改善の度合が比較的弱い．本方式が MAP の改善に貢献しないのは，この傾向によるところが大きい（MAP は検索結果上位の正解特許文書を持つ課題の順位変動に大きく左右される）と考えられる．

本評価実験から得られた各検索方式の有効性の検証結果を**表 3.7** にまとめる．ターム出現頻度を利用しないターム重み付け方式及び二段階検索方式は，評価データによって MAP の改善傾向が異なっているが，全体として見るとその効果は非常に高い．一方，尺度表現に着目したターム重み付け方式は，MAP の改善には貢献しないが正解特許文書の検索順位を全体的に押し上げるのに非常に有効な方式である．

なお，類似特許文書検索の出力例を**表 3.8** に示す．正解無効化特許 9 件のうち，8 件が検索結果上位 100 位以内に出力されており，比較的良い検索結果が得られた例である．

3.6 まとめ

本章では，代表的な特許請求項である【請求項 1】文章を入力として，その発明内容を無効化する特許を検索する類似特許文書検索の精度を向上させる方式を提案した．検索漏れ防止及び検索ノイズの低減の両方に対処すべく，特許文書の構成及び請求項文章の構造に着目した検索方式として，(1)検索段階に応じて検索タームの抽出範囲及び検索対象を変え，各検索段階における検索結果から最終的な類似度を算出する二段階検索方式，(2)請求項文章の記載に係る言語的特性を踏まえた，出現頻度を用いないクエリターム重み付け方式，(3)発明の特徴を表す尺度表現に着目したクエリターム重み付け方式を提案した．

これらの方式の有効性を検証すべく，約 170 万件の公開特許公報データに対して，NTCIR-4 特許検索タスクのフォーマルランの課題データ 103 件及び独自に用意した評価データ 228 件を使用した評価実験を行った．その結果，評価データセットによって傾向にばらつきがあるものの，ターム出現頻度を用いないクエリターム重み付け方式及び二段階検索方式は，全体として検索精度（平均精度）を向上させる効果が高いことが分かった．また，尺度表現に着目したクエリターム重み付け方式は，平均精度の改善には貢献しないものの，正解特許の正解順位を全体的に押し上げるのに有効な方式であることが分かった．

次章では，類似特許文書の検索精度を更に向上するために，クエリとその無効化特許の出願人に着目し，出願人の同一性が検索精度にどのように影響するかを定量的に分析するとともに，この分析結果に基づいて，複数の検索方式を組み合わせてより精度の高い検索方式を得る手法を提案する．

第４章　特許出願人に関する傾向分析とそれを適用した類似特許文書検索手法

4.1　はじめに

第 3 章で述べた類似特許文書検索において，クエリの発明内容を無効化する特許が，クエリと同一の発明者または出願人による特許である場合が少なくない．後述するように，クエリの発明内容を無効化する特許として特許庁審査官が引用した特許の約 21%が，クエリと同一の出願人による特許である．一方，1.3 節で述べたように，特許文書は記載項目が決まっており，文書構造が特許タグで規定されているが，その執筆スタイル（文章構成，構文，使用語彙）は執筆者または出願人によって異なっている．そして，この執筆スタイルの違いが類似特許文書検索の精度に少なからず影響していると考える．

しかし，これまでに提案された類似特許文書検索方式の有効性評価では，クエリとその無効化特許の出願人の同一性について考慮されてこなかった．「類似特許文書検索では出願人属性を利用することが有効である」という考え方は広く浸透している．しかし，類似特許文書検索において，出願人の同一性が検索精度に及ぼす影響を，詳細かつ定量的に評価した報告はなされていない．

そこで本章ではまず，クエリとその無効化特許の出願人が同じか違うかによって，類似特許文書検索の精度にどのような影響を与えるかについて，(1)文書属性，(2)使用タームの共通性，(3)検索の難易度という三つの観点から定量的に分析する[57][58]．次に，これらの分析結果を踏まえ，クエリとその無効化特許の出願人の同一性を利用した，類似特許文書検索方式の組み合わせ手法を提案する．すなわち，出願人の同一性の観点から個々の検索方式の有効性を評価し，その結果を踏まえて複数の検索方式を組み合わせて最適な検索方式を得る手法を提案する[58]．この手法は，新しい検索方式を提案するものではなく，既存の複数の検索方式をどのように組み合わせると，最も精度の高い検索方式となるかを，出願人の同一性を考慮して特定する手法である．

以下，4.2 節では，クエリとその無効化特許の出願人に関する傾向を，文書属性，使用タームの共通性，検索の難易度の観点から定量的に分析する．4.3 節では，第 3 章で提案した 4 種類の類似特許文書検索方式を採り上げ，出願人の同一性の観点からそれぞれの精度を評価し，精度の振る舞いを比較する．4.4 節では，上記傾向分析結果及び精度比較結果を反映させた検索方式の組み合わせ手法について提案し，その有効性を精度向上の観点から検証する．

4.2　出願人に関する傾向分析

4.2.1　文書属性の観点からの出願人傾向分析

ここでは，クエリとその無効化特許の出願人に係る傾向を，文書属性の観点から分析する．まず，出願人がクエリと同じ無効化特許の件数割合を分析した．1993 年から 1997 年に公開された特許の中で，同期間に公開された特許によって発明内容が無効とされたクエリ 210,755 件（対応する無効化特許は延べ 374,115 件）を分析対象とした．なお，無効化特許データは，独立行政法人工業所有権情報・研修館が発行している整理標準化データから抽出した．

傾向分析の結果を**表 4.1** に示す．無効化特許延べ 374,115 件のうち，78,807 件（21.1%）において，出願人がクエリと同じであった．また，出願人がクエリと同じ無効化特許を少なくとも一つ以上含むクエリは，210,755 件のうち 58,508 件（27.8%）であった．

次に，出願人がクエリと同じ無効化特許を含むクエリ 58,508 件を対象として，その技術分野のばらつきを分析した．技術分野として，クエリに付与されている筆頭の国際特許分類（IPC）のセクション（上位 1 桁，A-H の 8 分類）及びサブクラス（上位 4 桁，約 600 分類）を用いた．

出願人がクエリと同じ無効化特許を含むクエリの筆頭 IPC セクション別分布を**表 4.2** に示す．表 4.2 は，筆頭 IPC セクション（A-H）毎に，該当するクエリの総件数と，その中で出願人が同じ無効化特許を含むクエリの占める件数と，その割合を示している．出願人が同じ無効化特許を含むクエリの割合は，セクション D 及び C が 30%台で若干高いが，全セクションで割合が 20%を超えている．

また，IPC サブクラス別の分析では，210,755 件のクエリ集合に付与された筆頭サブクラスの異なり数 579 種類のうち，出願人が同じ無効化特許を含むクエリ集合 58,508 件に付与されたサブクラスの異なり数は，532 種類（91.9%）であった．これらの結果より，出願人がクエリと同じ無効化特許を含むクエリは，多くの技術分野に偏りなく存在していることが分かった．

更に，出願人がクエリと同じ無効化特許を含むクエリの出願人が，特定の出願人に偏っているか否かを分析した．分析結果を**表 4.3** に示す．本分析では，統計分析の精度を確保するために，クエリ 210,755 件の中で 10 件以上の出願件数を持つ出願人 1,664 組織によって出願された 202,393 件のみを分析対象とした．表 4.3 の第 1 欄は，ある出願人が出願した総クエリ件数に占める，出願人がクエリと同じ無効化特許を含むクエリ件数の割合を，

表 4.1　クエリとその無効化特許の出願人の同一性

#	項目		数値
1	クエリの件数		210,755 件
2	クエリを無効化する特許の延べ件数		374,115 件
3	出願人がクエリと同じ無効化特許	延べ件数	78,807 件
4		割合（#3/#2）	21.1%
5	出願人がクエリと同じ無効化特許を含むクエリ	件数	58,508 件
6		割合（#5/#1）	27.8%

表 4.2　クエリとその無効化特許の出願人の同一性（技術分野別）

技術分野（IPCセクション）	クエリ件数	出願人がクエリと同じ無効化特許を含むクエリ	
		件数	割合
A	14,476 件	3,975 件	27.5%
B	36,016 件	10,109 件	28.1%
C	22,500 件	8,476 件	37.7%
D	3,276 件	1,296 件	39.6%
E	7,874 件	1,860 件	23.6%
F	14,800 件	4,258 件	28.8%
G	58,020 件	15,295 件	26.4%
H	53,793 件	13,239 件	24.6%
合計	210,755 件	58,508 件	27.8%

表 4.3　クエリとその無効化特許の出願人の同一性（出願人別）

ある出願人のクエリ件数に占める，出願人がクエリと同じ無効化特許を含むクエリ件数の割合	左記に該当する出願人の異なり数（クエリ件数が 10 件以上を対象）		左記に該当する出願人によるクエリの総件数	
	異なり数	割合	総件数	割合
0- 10%	344	20.7%	12,214	6.0%
10- 20%	328	19.7%	26,159	12.9%
20- 30%	396	23.8%	58,803	29.1%
30- 40%	292	17.5%	61,943	30.6%
40- 50%	171	10.3%	36,559	18.1%
50%-	133	7.9%	6,715	3.3%
合計	1,664	100.0%	202,393	100.0%

6 段階の範囲に分けたものである．第 2 欄は，第 1 欄の各範囲に該当する出願人の数及びその割合を示している．第 3 欄は，第 2 欄に示された出願人によって出願されたクエリの合計件数及びその割合を示している．すなわち表 4.3 の第 1 行は，「ある出願人が出願した総クエリ件数に占める，出願人がクエリと同じ無効化特許を含むクエリ件数の割合が，0%から 10%の範囲にある出願人の数は，全体 1,664 組織の 20.7%にあたる 344 組織であり，これら 344 組織によって出願されたクエリの合計件数は，分析対象の全クエリ 202,393 件の 6.0%にあたる 12,214 件である」ことを示している．

ある出願人が出願した総クエリ件数に占める，出願人がクエリと同じ無効化特許を含むクエリ件数の割合（表 4.3 の第 1 欄）が，全体の平均値 27.8%（表 4.1）と比べて同等または少ない（30%以下である）出願人の数は 1,068 組織（344+328+396）であり，全体（1,664 組織）の 64.2%（20.7%+19.7%+23.8%）を占めた．一方，第 1 欄の割合が，全体の平均値 27.8%と比べて非常に高い（50%を超える）出願人も 133 組織（7.9%）存在している．しかし，これら 133 組織の出願人によって出願されたクエリの合計件数は 6,715 件（第 3 欄）であり，全クエリ 202,393 件の 3.3%と低い割合に留まった．これらの結果から，クエリと無効化特許の出願人が同じとなる現象は，特定の出願人に限定されていないことが分かった．

以上の分析結果より，クエリとその無効化特許の出願人が同じとなる現象は，無効化特許の延べ件数の 21%で起きており，どの技術分野や出願人にも見られる一般的な現象であることが分かった．

4.2.2 使用タームの共通性の観点からの出願人傾向分析

次に，クエリとその無効化特許の出願人に係る傾向を，使用するタームの共通性の観点から分析する．

多くの類似特許文書検索システムでは，文書中に使われているタームの共通性によって文書間の類似度を算出する方式を採用している．そこで，クエリとその無効化特許の間のターム使用傾向を分析した．すなわち，まず，クエリとその無効化特許との間に共通して使われるターム数の割合を算出した．そして，クエリとその無効化特許の出願人が同じか違うかによって，この割合値がどのくらい異なるかを比較した．

本分析では，4.2.1 節の分析で使用した特許データから，単独の出願人によって出願されたクエリと，その無効化特許のペア 20,000 組を抽出して分析対象とした．このうち 10,000 組は，クエリとその無効化特許の出願人が違う特許ペアで，残りの 10,000 組は，出願人が同じ特許ペアである．出願人が同じ特許ペア 10,000 組の内訳は，筆頭発明者も

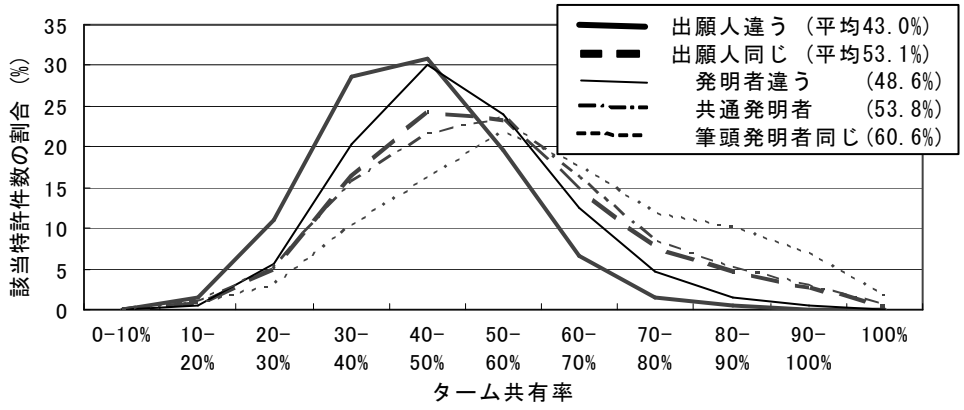


図 4.1　出願人の同一性とターム共有率の関係

同じ特許ペア 2,758 組，筆頭発明者は異なるが，共通の発明者が一人でも存在する特許ペア 2,285 組，共通の発明者がいない特許ペア 4,957 組である．

ターム抽出範囲は，発明の内容を端的に表すタームが多く含まれており，タームの使用傾向の違いが出やすいと考えられる請求項 1 のみとした．タームを抽出する形態素解析ツールとして茶筌を用い，茶筌で抽出されたタームすべてを分析対象とした．評価尺度として，「ターム共有率（クエリ中のタームの異なり数に占める，クエリにも無効化特許にも共通して出現するタームの異なり数の割合)」を用いた．クエリと無効化特許の出願人が同じか違うかによって，ターム共有率の平均値と，ターム共有率に対する特許ペア件数分布の傾向を比較した．なお，1 件のクエリに対する無効化特許が複数ある場合，共通するタームの割合は，無効化特許毎に別々に算出した．

ターム共有率に対する特許ペア件数の分布を**図 4.1** に示す．図の横軸は，10%刻みで範囲分けしたターム共有率を表し，縦軸は各範囲に該当する特許ペア件数を割合値で表している．また，図 4.1 の凡例の中に，ターム共有率の平均値を示している．出願人が同じ特許ペアでのターム共有率の平均は 53.1%で，出願人が違う特許ペアでの平均 43.0%に比べ，10.1 ポイント高い．すなわち，出願人が同じ場合，共通のタームが使われる傾向が強いと言える．また，「筆頭発明者も同じ」場合のターム共有率の平均が 60.6%と最も高いが，「出願人は同じだが発明者が全く違う」場合でも 48.6%であり，出願人が違う場合の平均 43.0%に比べ，5.6 ポイント高い．更に，全体の件数分布を見ても，出願人が違う場合の件数分布（太い実線）に比べ，出願人が同じ場合（太い点線，細い実線，1 点鎖線，破線）の方が，どれも分布全体が右方向（ターム共有率の高い方向）にずれており，共通するターム

が使われる割合が高くなっている．この割合の違いは，次節で述べる無効化特許検索の難易度の差異を説明する根拠の一つにもなっている．

次に，出願人別のターム使用傾向を分析した．すなわち，特定の技術分野において，タームが出現する特許文書数の割合が，出願人によってどの程度偏っているかについて比較した．

2002 年に公開された特許 374,550 件に含まれる「検索技術」に関する（公開特許公報に記載されるテーマが 5B075「検索装置」である）特許 5,330 件のうち，出願件数の多い上位 10 出願人による特許 1,925 件を分析対象とした．ターム抽出範囲は請求項 1 の文章のみとし，茶筌によってタームを自動抽出した．出願人別ターム使用傾向の評価尺度として，分析対象となる文書数に対する各タームの出現文書数の割合を出願人別に算出し，その値を 10 出願人で比較した．そして出願人別の割合の最大値と最小値の差（以下，「割合格差」

表 4.4　「検索技術」に関する特許文書における出願人別のターム使用傾向の比較

	特許文書総数	5B075 特許文書数	特許出願件数上位10社											
			合計	A	B	C	D	E	F	G	H	I	J	平均
特許文書数	374550件	5330件	1925件	262件	218件	209件	209件	191件	190件	188件	177件	141件	140件	193件
発明者異なり数	－	4421人	1529人	180人	196人	194人	119人	150人	149人	161人	127人	129人	124人	153人
ターム見出し	そのタームが使用されている特許文書数の割合 (%)													割合格差
(スペース)	81.3	80.0	83.4	98.9	92.7	**44.0**	92.8	79.1	69.0	85.6	98.9	72.3	98.6	54.9
文書	0.3	8.4	14.7	8.4	7.3	14.4	53.6	13.6	**0.0**	10.1	13.6	14.9	9.3	53.6
方法	17.4	27.4	29.4	64.5	21.6	50.2	18.2	34.0	**11.1**	11.7	20.3	29.8	15.0	53.5
備える	35.3	45.0	41.8	**17.9**	41.3	29.7	50.2	24.6	47.4	57.5	59.3	41.8	65.0	47.1
手段	25.6	52.8	54.4	**33.2**	52.8	41.2	55.0	53.9	73.7	64.9	79.1	45.4	54.3	45.9
者	4.5	24.8	22.7	49.6	32.6	25.8	12.4	14.7	13.2	22.9	**6.2**	19.9	15.0	43.4
この	14.4	12.0	12.4	17.2	13.8	3.4	12.0	42.9	7.9	5.9	**0.0**	0.7	16.4	42.9
装置	37.1	40.3	47.8	**30.5**	33.0	34.5	53.1	38.2	72.1	62.2	63.3	48.2	55.7	41.6
システム	8.7	45.6	38.0	30.9	59.6	46.4	45.5	35.6	30.0	34.0	33.3	**18.4**	39.3	41.2
前記	60.5	76.0	77.7	78.2	91.7	66.0	74.6	90.1	66.3	88.3	89.3	73.1	**51.4**	40.3
上記	9.7	9.3	11.2	9.2	**1.4**	11.5	2.4	6.3	27.4	5.9	6.2	12.1	40.7	39.3
端末	4.3	29.9	24.6	34.4	48.2	14.8	**12.9**	23.0	22.6	23.4	14.7	14.2	30.7	35.3
具備	5.6	8.3	8.7	3.8	6.4	2.4	3.8	36.1	12.1	15.4	3.4	**1.4**	**1.4**	34.7
利用	3.0	17.9	18.4	40.5	26.2	21.1	12.0	14.1	11.1	17.0	**6.8**	11.4	10.7	33.7
該	24.9	22.9	24.3	37.4	23.4	30.1	30.6	13.6	13.2	**6.9**	38.4	28.4	13.6	31.5
において	38.7	29.8	34.0	37.0	28.9	47.4	43.1	33.0	31.1	**16.5**	29.4	36.9	34.3	30.9
ある	39.3	47.2	45.7	64.5	49.5	39.7	36.8	**34.0**	42.1	45.7	58.8	42.6	34.3	30.5
特徴	82.6	80.7	84.9	82.1	88.5	90.9	90.9	91.6	82.6	**68.6**	97.2	76.6	75.7	28.6
記憶	5.8	27.8	24.7	**11.8**	22.0	20.6	23.9	30.9	33.2	31.9	23.7	17.0	40.0	28.2
提供	2.3	25.1	23.0	37.0	30.3	18.7	**9.6**	18.3	33.2	18.6	12.4	25.5	20.7	27.5
有する	25.0	23.7	28.2	40.1	28.9	39.7	23.4	**13.1**	21.1	23.4	36.7	25.5	22.9	27.0
だ	48.4	54.7	52.8	69.1	55.5	47.4	46.4	45.0	52.6	52.1	61.6	**42.6**	46.4	26.5
として	10.5	20.0	23.0	37.8	21.6	21.1	29.2	24.1	**12.1**	15.4	19.8	27.0	15.0	25.7
こと	84.5	83.5	86.1	84.0	90.8	89.0	91.4	93.2	82.6	**71.8**	97.2	80.9	76.4	25.4
で	39.9	35.2	35.6	46.2	34.9	41.2	40.7	36.7	**21.1**	29.3	39.6	28.4	30.7	25.1
接続	12.7	22.3	18.1	17.2	31.2	17.7	17.2	14.1	17.4	16.5	17.5	**6.4**	22.1	24.8
介す	10.1	23.0	19.4	22.9	32.6	11.5	16.8	17.3	26.8	14.9	18.1	**7.8**	20.7	24.8
情報	11.7	71.1	72.0	78.2	76.2	75.1	**55.0**	67.0	79.5	76.6	65.5	70.9	73.6	24.5
コンテンツ	0.6	7.9	10.0	19.9	7.3	6.2	3.8	**2.6**	26.8	9.6	6.8	6.4	5.7	24.2
送信	5.2	23.2	22.0	26.7	33.9	16.8	13.4	**11.0**	31.1	22.3	18.6	19.2	24.3	23.0
検索	1.2	39.4	39.3	40.1	50.9	36.4	37.3	42.9	**28.4**	31.9	42.9	34.8	47.1	22.5
が	59.1	60.8	60.3	72.1	70.2	59.8	56.0	55.0	51.1	53.7	61.0	**49.7**	67.9	22.5
情報処理	0.8	5.0	7.3	**1.9**	2.8	5.7	6.2	2.6	24.2	5.3	13.6	7.8	6.4	22.3
データベース	1.5	28.5	25.3	30.9	26.2	33.0	19.6	28.3	**14.7**	20.7	17.5	25.5	36.4	21.7
により	20.9	24.2	25.1	**17.6**	19.7	21.5	22.0	31.9	39.0	22.9	32.8	20.6	27.9	21.4
から	44.1	66.5	67.0	75.2	69.3	63.6	67.5	63.9	**56.3**	66.5	62.7	68.1	76.4	20.1

注1：網掛けされたターム見出しは一般語とみなされるもの
注2：網掛けされた数値はそのタームにおける最大値を示し、太字の数値は最小値を示す。

と呼ぶ）が大きいタームほど出願人別ターム使用傾向に偏りがあると判定した.

割合格差が 20 ポイント以上であるタームの使用傾向を**表 4.4** に示す（一部抜粋）. 割合格差の大きいタームの中には,「コンテンツ」「検索」「データベース」など, 検索分野でよく使われるタームのほかに, 発明内容に関係しない一般語（表 4.4 でターム見出しに網掛けされたターム）も含まれている.

タームを個別に見ると, ターム「文書」の割合格差は 53.6 ポイント（53.6%(D 社)-0.0%（F 社)）と大きい. この大きな格差は, 今回分析範囲とした技術分野が情報検索全般を網羅しており, F 社では文書を対象とした検索に関する特許を出願していないために生じたと説明できる. しかし,「この(割合格差 42.9 ポイント)」「該(同 31.5)」「ある(同 30.5)」などは発明内容の観点からは格差の理由が説明できない. また, ほぼ同じ意味であるターム「前記」「上記」を見ると, B 社では「前記」を専ら使用している（91.7%）のに対して, J 社では「上記」を使用している特許の割合が非常に高い（40.7%）. これらのターム使用傾向の偏りは, 出願人毎の執筆スタイルの嗜好性の違いによるものと言える.

4.2.3 検索の難易度の観点からの出願人傾向分析

更に, クエリとその無効化特許の出願人に係る傾向を, 類似特許文書検索の難易度の観点から分析する.

4.2.2 節で述べたように, クエリとその無効化特許の出願人が同じ場合, ターム共有率が高いという傾向がある. このことから,「出願人がクエリと同じ無効化特許の検索は, 出願人がクエリと違う無効化特許の検索よりも容易である」と推測できる. そこで, この推測が正しいかを実データを用いた精度評価実験により検証する.

本実験では, NTCIR-5 特許検索タスクのフォーマルラン課題のクエリ請求項 1,189 件を入力文章とし, 1993 年から 2002 年に公開された特許約 350 万件を検索対象とした. 正解特許は, 同タスクで用意された無効化特許延べ 2,065 件とした（この中には出願人がクエリと同じ特許 341 件（16.5%）を含む）. 検索エンジンは GETA を用い, 形態素解析ツールは茶筌を用いた. 評価尺度として平均精度（式(3.4)参照）と,「ある検索出力順位以内に出力された正解無効化特許件数の割合」を用いた.

実験結果を**表 4.5** に示す. 表 4.5 では正解無効化特許を, 全体(#1, 正解件数 2,065 件), 出願人がクエリと同じ場合（#2, 同 341 件), 出願人がクエリと違う場合（#3, 同 1,724 件）に分けて精度評価している. 検索結果をクエリ毎に 1,000 件ずつ出力した場合に, 全体（#1）での平均精度が 0.1496 であった. これに対し, 出願人がクエリと同じ場合（#2）

表 4.5　出願人の同一性と類似特許文書検索の難易度の関係

		全体 #1 （正解 2,065 件）	出願人同じ #2 （正解 341 件）	出願人違う #3 （正解 1,724 件）
平均精度		0.1496	0.3478	0.1029
		含まれる 正解件数の割合	含まれる 正解件数の割合	含まれる 正解件数の割合
検索結果 出力順位	1	7.3%	21.1%	4.5%
	1- 10	20.7%	49.0%	15.1%
	1- 50	36.4%	64.2%	30.9%
	1- 100	43.6%	71.3%	38.2%
	1- 300	58.2%	83.9%	53.1%
	1- 500	64.4%	87.4%	59.9%
	1-1000	72.7%	92.1%	68.9%

表 4.6　出願人の同一性と類似特許文書検索の難易度の関係（出願人でフィルタリング）

		出願人同じ #2 （正解 341 件）	出願人同じ #4 クエリの出願人で フィルタリング （正解 341 件）
平均精度		0.3478	0.4477
		含まれる 正解件数の割合	含まれる 正解件数の割合
検索結果 出力順位	1	21.1%	27.3%
	1- 10	49.0%	62.5%
	1- 50	64.2%	81.2%
	1- 100	71.3%	86.8%
	1- 300	83.9%	91.5%
	1- 500	87.4%	92.1%
	1-1000	92.1%	92.1%

表 4.7　検索結果上位に占める出願人が同じ特許文書件数の割合

検索順位（N）	1000 位 以内	500 位 以内	100 位 以内	50 位 以内	10 位 以内
検索結果上位Ｎ位以内に含まれる，出願人がクエリと同じ無効化特許件数の割合	6.2%	7.6%	12.2%	14.7%	21.1%

の平均精度は 0.3478 と格段に高く，逆に出願人が違う場合（#3）の平均精度は 0.1029 と低い．また，上位 100 位以内に出力される正解無効化特許の件数割合は，全体（#1）で 43.6%であるのに対し，出願人がクエリと同じ場合（#2）は 71.3%，出願人がクエリと違う場合（#3）は 38.2%と大きな格差が生じている．この傾向はどの出力順位でも見られる．

更に，出願人がクエリと同じ場合（#2）において，検索結果上位 1,000 件からクエリと同じ出願人の特許だけをフィルタリングして，結果出力した時の評価尺度を計算した結果を**表 4.6** に示す．平均精度は，0.3478 から 0.4477 に，上位 100 位以内に出力される正解無効化特許の件数割合は，71.3%から 86.8%に大幅に向上する．フィルタリング対象を検索結果上位 1,000 件に限定しなければこれらの値は更に良くなると予想される[10]．なお，本フィルタリングにより抽出されるクエリと同じ出願人の特許件数の平均は，62 件（最多 765 件，最少 0 件）であった．

以上の結果より，出願人がクエリと同じ無効化特許の検索は，出願人が違う場合に比べて容易であること，全体の平均精度は，出願人がクエリと同じ無効化特許（全体の 16.5%（341 件/2065 件）を占めるに過ぎない）の平均精度の影響を強く受けていることが分かった．

ところで，上記評価実験結果と，4.2.2 節で得られた「出願人が同じ特許ペアのターム共有率は比較的高い」という傾向から，「クエリと出願人が同じ特許は検索結果の上位に出力されやすい」と推測できる．この推測が正しい場合，出願人がクエリと同じ特許が，検索結果の上位に集まりやすくなり，出願人が違う無効化特許が検索結果の下位に埋もれてしまいやすいと考えられる．そこで，上記推測が正しいかを検証するために，出願人がクエリと同じ特許（正解無効化特許以外の特許も含む）が，検索結果の上位にどのくらいの件数割合で出力されているかを分析した．

分析結果を**表 4.7** に示す．表 4.7 は，検索結果上位 N 位（N=1000,500,100,50,10）以内に含まれる，出願人がクエリと同じ特許の件数割合を示している．検索結果上位 1,000 位以内での割合は 6.2%と低いが，100 位以内で 12.2%，10 位以内で 21.1%であり，検索結果の上位ほど出願人がクエリと同じ特許の密度が高い（集まりやすい）傾向がある．この結果から，出願人がクエリと違う無効化特許の検索が難しい要因の一つとして，出願人がクエリと同じ特許が検索結果の上位に多く出力されていることが挙げられる．

---

[10] 本フィルタリングでは，検索結果からクエリと同じ出願人の特許を抜き出している．このとき，検索結果を何件出力すればフィルタリング後の件数が 1,000 件になるかは未知である．そこで本評価では，フィルタリング前の検索結果を予め 1,000 件に限定している．

4.2.4　出願人が同じとなる現象の要因分析

これまでの分析結果から，クエリとその無効化特許の出願人が同じとなる現象が，技術分野や出願人を問わず高頻度で起きていることが分かった．ここでは，この現象が起きる要因として以下の 2 点を挙げ，定性的に考察する．

(1) 改良特許の新規性/進歩性の拒絶

既出願特許 A の発明の一部を改良し，その改良部分に新規性または進歩性を持たせて，別の特許 A1 として出願することがしばしばある．しかし，改良部分を無効化する特許 B が存在する場合，特許 A1 は，その母体となる特許 A と特許 B の組み合わせによって実現可能となり，この二つの無効化特許によって拒絶される．この場合，特許 A と A1 は同一の出願人である場合がほとんどであるため，クエリとその無効化特許の出願人が同じとなる現象が起きる．

(2) 特許情報共有の不備

特許出願件数の多い大企業では，類似する研究が別の部署で独立に行われることがしばしばある．この場合，ある発明内容が既に他の部署で特許出願されていることを知らずに，同一の内容で出願すると，この特許は同一出願人である別部署から出願された特許によって拒絶される．

これまでに述べた出願人傾向分析においても，上記要因を考慮すべきであったが，そのためには出願人の組織情報などが必要となる．しかし，その入手が困難なため，本研究ではこれらの要因を考慮していない．

4.2.5　出願人傾向分析結果から得られる技術課題

以上の分析結果から，出願人がクエリと同じ無効化特許の検索は容易であり，出願人がクエリと違う場合の検索は難しいことが分かった．このことから，類似特許文書検索の精度向上に係る研究では，「出願人が違う無効化特許の検索精度をいかに向上させるか」という技術課題を解決することが重要であると考える．

類似特許文書検索の精度向上に関する従来研究では，検索方式の有効性を評価する際に，出願人の同一性について考慮していない．すなわち，出願人がクエリと同じ無効化特許と，出願人がクエリと違う無効化特許が混在した評価データセットを用いて一括評価している．しかし，「出願人が違う無効化特許の検索精度をいかに向上させるか」という上記技術課題を解決するためには，クエリとその無効化特許の出願人が同じか否かによって，検

索方式の有効性がどのように変化するかを別々に評価し，その検索方式が，出願人の違う無効化特許の検索精度の向上に貢献しているかを検証することが不可欠と考える．

そこで次節では，これまでに提案されてきた種々の類似特許文書検索方式のうち，第 3 章で提案した 4 種類の検索方式を採り上げ，その検索精度を出願人の同一性の観点から評価する．すなわち，無効化特許の出願人がクエリと同じか否かによって，各検索方式の精度的振る舞いがどのように異なるかを検証する．

4.3 出願人の同一性から見た類似特許文書検索方式の精度的振る舞いの検証

4.3.1 検証に用いる検索方式の概要

本検証では，第 3 章で述べた以下の 4 種類の検索方式を採り上げる．そして，各検索方式の精度的振る舞いが，出願人の同一性によってどのように異なるかを定量的に検証する．

(1) ターム出現頻度の固定（3.3.1 節参照．以下，「TF1」と呼ぶ．）

クエリから抽出されるタームの出現頻度（Term Frequency，TF）をすべて 1 に固定してタームの重みを算出する（式(3.3)参照）．

(2) 不要語除去（3.4 節参照）

ここでは，検索対象となる特許文書集合のうち，20%以上の特許の請求項に出現する単語（31 語）をクエリタームから除去する．

(3) 二段階検索（3.2 節参照）

第一段階では，特許文書全文を対象として広く浅く検索し，検索結果上位 1,000 件に絞る．第二段階では，請求項のみを対象として狭く深く検索し，第一段階のスコアに第二段階のスコアの一部をマージする．

(4) 尺度表現利用（3.3.2 節参照）

請求項にしばしば現れる物性を表すターム 361 語（「速度」「膨張率」など）の重みを高くする．

上記 4 種類の検索方式は，類似特許文書検索精度を向上するための一般的アプローチである，「検索に使用するタームの抽出方法」（上記方式(2)に対応），「タームの重要度を示

す重みの算出方法」(上記方式(1)(4)に対応),「類似度の算出方法」(上記方式(3)に対応)にそれぞれ対応しており,決して特殊な検索方式ではなく,一般的な方式であると言える.

4.3.2 検証結果と考察

各検索方式を適用しない場合と適用した場合で,検索精度がどのように異なるかについて,出願人の同一性の観点から実験評価する.本実験では,4.2.3 節で使用した NTCIR-5 特許検索タスクの実験環境(クエリ 1,189 件,検索対象特許 350 万件)を用いた.また,評価尺度として平均精度(MAP)を用いた.

実験結果を**表 4.8** に示す.表 4.8 の#0 は,検索精度の比較基準となる方式(ベースライン)であり,上記 4 方式のどれも適用しない場合の平均精度である.#1 から#4 は,#0 に上記各方式を適用した場合の平均精度である.

出願人が同じか否かを区別せずに,正解無効化特許全体で算出した時の平均精度(表 4.8 の「全体」)が比較基準(#0)と比べて向上している検索方式は,二段階検索(#3)のみである(0.1496⇒0.1550).しかし,出願人が同じか違うかによって,正解無効化特許を二つに分けて評価すると,傾向が変わる.すなわち,TF1(#1)では,全体で見ると平均精度は悪化している(0.1496⇒0.1492)が,出願人がクエリと同じ場合の平均精度はやや改善傾向にある(0.3478⇒0.3497).不要語除去(#2)では,出願人がクエリと同じ場合の平均精度は悪化している(0.3478⇒0.3425)が,出願人がクエリと違う場合の平均精度はやや改善している(0.1029⇒0.1034).すなわち,TF1 と不要語除去は,出願人の同一性という観点から見ると,全く逆の精度的振る舞いをする検索方式であることが分かる.また,二段階検索(#3)では,出願人が同じ場合も違う場合も平均精度は改善されているが,出願人が違う場合の方が改善の度合が高い.更に,尺度表現利用(#4)では,不要語除去(#2)と同様に,出願人が違う場合において精度改善傾向が見られる.このように,全体

表 4.8 出願人の同一性による 4 種類の検索方式の精度的振る舞い

実験ＩＤ		#0	#1	#2	#3	#4
検索方式		ベース	ＴＦ１	不要語除去	二段階検索	尺度表現
平均精度	全体 (正解特許 2,065 件)	0.1496	0.1492 (−0.3%)	0.1485 (−0.7%)	0.1550 (+3.6%)	0.1480 (−1.1%)
	出願人が同じ (正解特許　341 件)	0.3478	0.3497 (+0.5%)	0.3425 (−1.5%)	0.3501 (+0.7%)	0.3374 (−2.9%)
	出願人が違う (正解特許 1,724 件)	0.1029	0.1014 (−1.5%)	0.1034 (+0.5%)	0.1090 (+5.9%)	0.1031 (+0.2%)

で見た時の各検索方式の精度的振る舞いと，出願人が同じか違うかの観点から見た時の精度的振る舞いに違いが生じていることが分かる．

この原因の一つとして，クエリとその無効化特許で使用されるタームの共通性が挙げられる．以下では，不要語除去の精度的振る舞いを例にとって考察してみる．一般の類似特許文書検索では，検索に使われるタームの一致度によって類似度を算出している．出願人が同じ特許では，発明内容に直接関係しない一般語の使われ方が類似する（共通して使われる確率が高い）傾向にあることが，表 4.4 の結果から分かっている．すなわち，この一般語の存在が，類似度の値に多少なりとも貢献していると考えられる．一方，不要語除去は，これら一般語を検索に貢献しない不要語とみなしてタームから除去する処理である．したがって，一般語が除去されるほど類似度が下がり，これらの一般語が除去されない（類似度が下がらない）他の特許に比べ，検索順位が相対的に低下すると考えられる．

4.4 出願人の同一性を考慮した類似特許文書検索方式の組み合わせ手法

4.4.1 組み合わせ手法の提案

一般に，一つの文書検索システムにおける検索アルゴリズムは，複数の検索方式を組み合わせたものとなっている．すなわち，種々の検索方式のうち，精度向上に貢献する方式を組み合わせることによって，文書検索システム全体としての検索精度を更に向上させている．この際，どの検索方式をどのように組み合わせると，検索精度が最も良くなるのかを検証する評価作業が不可欠となる．

従来手法では，最も検索精度が高くなる検索方式の組み合わせを特定する際にも，クエリとその無効化特許の出願人の同一性について考慮していない．すなわち，出願人がクエリと同じ無効化特許と，クエリと違う無効化特許を区別せずに一括評価していた．

これに対して本研究では，4.2.5 節で提起した技術課題「出願人が違う無効化特許の検索精度をいかに向上させるか」に対処する一手法として，「出願人の同一性を考慮した，類似特許文書検索方式の組み合わせ手法」を提案する．この組み合わせ手法は，4.2 節で述べた出願人に関する傾向である，(1)出願人がクエリと同じ無効化特許の検索精度が全体の検索精度に大きく影響していること，(2)検索精度を向上させる検索方式の精度的振る舞いは，無効化特許がクエリと同じ出願人であるか否かによって異なること，の 2 点を考慮した手法である．すなわち，本手法では，個々の検索方式を複数組み合わせた時の検索精度を評価する際に，クエリとその無効化特許の出願人が違う場合の検索精度の振る舞

いに着目する．そして，その検索精度がより高くなるように検索方式を組み合わせるという手法である．

4.4.2　組み合わせ手法の有効性検証

前節で提案した検索方式の組み合わせ手法の有効性について，4.3 節で精度評価した 4 種類の検索方式を用いて検証する．4 種類の検索方式を組み合わせた時に得られる検索精度を**表 4.9** に示す．#0 は比較基準であり，#1-#4 は各検索方式を単独で適用した場合（表 4.8 と同一），#5-#10 が複数の検索方式を組み合わせた場合である（明らかに精度が低い組み合わせについては表 4.9 から除外している）．

出願人の同一性を考慮しない従来の組み合わせ手法では，無効化特許全体（表 4.9 の「全体」）での検索精度のみを比較することによって，検索精度が最も高い組み合わせを特定する．その結果として，TF1 と二段階検索の 2 方式を組み合わせた検索方式#6（平均精度 0.1572）が選ばれる．これに対して，本論文で提案する組み合わせ手法では，出願人がクエリと違う無効化特許の検索精度の向上を重視する．そのため，出願人がクエリと異なる無効化特許（表 4.9 の「出願人違う」）での検索精度が最も高い組み合わせに着目する．その結果として，#6 とは検索方式の組み合わせの異なる，不要語除去，二段階検索，尺度表現利用の 3 方式を適用した#9（平均精度 0.1101）が選ばれる．#9 を構成する 3 つの検索方式は，表 4.8 で示したように，出願人がクエリと違う無効化特許の検索精度において改善傾向が見られたものであり，検索方式を組み合わせた場合でもこの精度的振る舞いが反映されている．

表 4.9　検索方式を組み合わせた場合の検索精度の比較

実験 ID		#0	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
検索方式	TF1	－	○	－	－	－	○	○	－	○	－	○
	不要語	－	－	○	－	－	○	－	○	○	○	○
	二段階	－	－	－	○	－	－	○	○	○	○	○
	尺度表現	－	－	－	－	○	－	－		－	○	○
平均精度	全体	.1496	0.1492 −0.3%	0.1485 −0.7%	0.1550 +3.6%	0.1480 −1.1%	0.1483 −0.9%	0.1572 +5.1%	0.1553 +3.8%	0.1555 +3.9%	0.1549 +3.5%	0.1541 +3.0%
	出願人同じ	.3478	0.3497 +0.5%	0.3425 −1.5%	0.3501 +0.7%	0.3374 −2.9%	0.3462 −0.5%	0.3627 +4.3%	0.3483 +0.1%	0.3539 +1.8%	0.3445 −0.9%	0.3544 +1.9%
	出願人違う	.1029	0.1014 −1.5%	0.1034 +0.5%	0.1090 +5.9%	0.1031 +0.2%	0.1015 −0.2%	0.1086 +5.5%	0.1099 +6.8%	0.1081 +5.1%	0.1101 +7.0%	0.1065 +3.5%

従来手法で得られた組み合わせ（#6）と，本論文で提案する手法で得られた組み合わせ（#9）の検索精度（平均精度）を比較してみる．全体では#6 の方が 1.5%高い（0.1549⇒0.1572）が，出願人がクエリと違う無効化特許のみの場合では，逆に#9 の方が 1.4%高い（0.1086⇒0.1101）．本提案手法では，全体の検索精度が低下している（0.1572⇒0.1549）が，これは出願人がクエリと同じ無効化特許に係る精度低下（0.3627⇒0.3445）に起因している．すなわち，出願人が違う無効化特許の検索精度を上げようとすると，出願人が同じ無効化特許の検索精度が下がるというトレードオフの傾向が見られる．これは，表 4.8 で示したように，不要語除去及び尺度表現利用の精度振る舞いにトレードオフの傾向があることによる．出願人がクエリと同じ無効化特許は，表 4.6 に示したように，クエリの出願人で検索結果をフィルタリングするなどによって比較的容易に検索できることから，この精度低下はある程度までは容認できると考える．

また，#6 と#9 において，出願人が違う無効化特許の検索精度の差は 1.4%（(0.1101-0.1086）/0.1086）と低い．これは，この差に貢献している検索方式である，不要語除去及び尺度表現利用による精度改善の度合が小さいためと考えられる．不要語除去については 31 語の不要語しか使用していないが，更に不要語を追加登録することにより，検索精度改善の度合が大きくなる[57]．したがって，不要語の拡充により，本組み合わせ手法によるさらなる精度向上が期待できる．また，今回の評価で採り上げていない検索方式を用いた組み合わせによって，検索精度を更に向上させることも可能と考える．

本手法において出願人の同一性を考慮すると，検索精度が最良となる検索方式の組み合わせが変わることがあることを実験により検証した．平均精度自体は 0.1 前後であり，無効化特許検索の実用性の観点からはさらなる精度向上が必要である．しかし，文書検索において精度が劇的に向上する方式が現れるというのは考えにくく，地道な方式改善の積み重ねが必要である．本論文で述べた出願人の同一性を考慮した検索方式の組み合わせ手法はその一つとして有効であると考える．

4.5 まとめ

無効化特許の検索精度を向上するために，クエリとその無効化特許の出願人の同一性に着目して検索方式を組み合わせる手法について述べた．本手法を導くために，出願人に関する傾向を分析した．その結果，(1)クエリとその無効化特許の出願人が同じであるという現象が技術分野を問わず高頻度で起こっている，(2)出願人が同じ無効化特許の検索は比較的容易である，(3)これまでに提案されてきた検索方式の精度的振る舞いが出願人の同一性によって異なる，という傾向があることを定量的に示した．また，本組み合わせ手

法が有効であることを，NTCIR-5 特許検索タスクの評価データを用いて検証した．

今後の課題として 3 点が挙げられる．第一に，本章で扱わなかった検索方式について精度振る舞いを検証することが挙げられる．異表記展開や同義語展開，質問文拡張，適合性フィードバックなどの検索方式に適用することによって，本研究で提案した組み合わせ手法の有効性が増すと考えている．例えば，異表記展開では，同じ出願人が執筆する特許においては，使用するタームの表記がある程度統一されていることを考えると，不要語除去と同じく，出願人の同一性によって精度振る舞いが大きく変わると予想される．

第二に，本組み合わせ手法と評価尺度との関係について，更に検証していくことが挙げられる．検索方式の最適な組み合わせは，検索精度の評価で用いる評価尺度の性質も影響すると考えている．本研究では平均精度を採用したが，平均精度の値は，出願人がクエリと同じ無効化特許(検索結果上位に出力される無効化特許)の影響を大きく受ける．逆に，評価尺度として例えば，「ある検索方式を適用した時に，検索順位が改善された無効化特許件数の割合」で評価すると，出願人が違う無効化特許の方が件数としては大きいため，上記割合値は，出願人の違う無効化特許の振る舞いの影響を受けやすくなり，表 4.9 に示したような差異が見られない可能性もある．したがって，本章で提案している組み合わせ手法と評価尺度との関係についても今後検証していく必要があろう．

そして第三に，出願人が同じとなる場合のケースを細分類して，それぞれのケース毎の要因を分析して検索精度の向上に反映させることが挙げられる．

第５章　従来研究の動向と本研究との比較

本章では，文書検索技術及び分類付与技術に関するこれまでの研究動向について概観し，本研究と比較する．5.1 節では文書検索の歴史的経緯を概観する．5.2 節では文書検索技術の研究成果として，これまでに提案されてきた各種の検索モデル，インデクシング方式，ターム抽出及び重み付け方式，検索アルゴリズムについて述べる．5.3 節では分類付与技術に係る研究動向について概観する．5.4 節では特に特許文書を対象とした類似特許文書検索及び分類自動付与技術に関する研究動向について述べ，本研究のアプローチと比較する．

5.1　文書検索の歴史的経緯

「情報検索」という言葉には，広義の定義と狭義の定義が存在する[108]．広義の情報検索とは，「ユーザの持つ問題を解決できる情報を見つけること」であり，狭義の情報検索とは，「ユーザの検索質問に適合する文書を文書集合の中から見つけること」である．一般に情報検索という場合，狭義の定義として使われることがほとんどである．また，Taylor は情報検索におけるユーザの情報要求のレベルをその具体化の度合によって 4 段階に分類している[107]が，本研究を含め多くの情報検索システムでは，情報要求が最も具体化されている状態，すなわちユーザが所望の情報の内容を形式化できる状態（compromised need，調整済みの要求[108]）にあることを前提としている．各情報はテキスト，静止画，動画，音声などから構成されるが，これらは「文書」という単位でまとめられることがほとんどであることから，「文書検索」という言葉を使うことが多く，本論文でも文書検索という言葉を用いている．

さて，文書検索の研究は半世紀前の 1950 年代から米国などの英語圏を中心として始まった．当初は出現頻度などの統計情報をベースとして，索引語として文章を特徴付けるタームの抽出技法及びその応用に関する研究が主流であった[50][51][52]．1970 年代から 1980 年代にかけて，ベクトル空間モデル[96]や確率モデル[86]，TF-IDF 法[88]といった，現在の文書検索アルゴリズムの基礎となる検索モデルやタームインデクシング方式が提案されるようになり，SMART[88][89]に代表される高精度な文書検索システムが開発されるようになった．

一方，日本における情報検索の研究は，1970 年代から盛んになった[16]．欧米諸国と比較すると，日本における情報検索の発展は日本語文章入力技術の発展に大きく関わってい

る.1978 年における漢字コード JIS-1978 の標準化や 1980 年代のワードプロセッサの普及は，情報検索の発展の加速化に大きく貢献した．1980 年代に入ると，「オフィスオートメーション」のキーワードのもと，企業・大学内の文書が仮名漢字文章で電子化され，蓄積されるようになってきた．またこれらの技術動向に伴い，1978 年に開始した特許情報オンライン検索サービスや，1981 年に稼動した科学技術文献検索サービス JOIS-Ⅱに代表されるオンライン情報検索サービスも普及してきた．しかし，当時の情報検索サービスは情報部門の専門家向けであり，公衆電話回線による通信によって提供されるものであり，学術文献や特許文書の書誌情報を検索したり，予め人手で収集された統制語から検索タームを選択して検索したりする機能に留まっていた．しかし，1990 年代に入ると文書全文を自由なタームで検索可能とする全文検索技術が実用化され，その後も種々の特徴を持つ検索システムが次々に提供されるようになった．

日本，諸外国を問わず，1980 年代までの文書検索では，検索対象は学術論文や図書情報などの特定の情報に限られており，情報の提供者も限られていた．しかし，1990 年代に入ってインターネットとしてネットワークインフラが整備されると状況は一変した．検索の対象となる情報のマルチメディア化，情報量の爆発的増加，情報発信者の多様化による文書品質の多様化，情報の更新頻度の増加など，それまでとは全く違った状況になり，それに伴って情報検索に対するユーザ要求も多種多様化していった．このような社会的基盤の劇的な変化が，情報検索の研究分野を活性化させる契機となった．特に，自然言語処理及び統計処理を情報検索の分野に適用することにより，情報検索技術は更に発展した．

一方，ハードウェア性能の向上により，性能面でも情報検索技術は飛躍的に進歩した．主要な Web 検索エンジン[29][113]に代表されるように，億単位の文書情報から所望の情報を数秒以内で検索できるようになっている．更に現在では，地域や技術分野，文書種別，利用目的などに特化した Web 検索サービスも数多く提供されており[91][100]，情報検索は我々の生活に欠かせないツールとなっている．

1980 年代までの文書処理研究では，提案した方式の有効性を評価する環境の構築が研究推進上の大きな障壁の一つであった．しかし，1980 年代後半から活発化したコーパスベース研究のためのテキストコーパスの整備，1992 年に米国で始まった TREC (Text REtrieval Conference) [11]や，2000 年に欧州で始まった CLEF (Cross Language Evaluation Forum) [12]，1996 年に公開された日本語情報検索システム評価用テストコレクション BMIR-J1 (BenchMark for Information Retrieval systems for Japanese texts ver.1) [40]，1998

[11] TREC　http://trec.nist.gov/

[12] CREF　http://clef.iei.pi.cnr.it/

年に国立情報学研究所などが主催して始まった情報検索システム評価用テストコレクション構築プロジェクトNTCIR（NII-NACSIS Test Collection for IR Systems）[13]などにおいて，大量の文書コーパス及び評価データセットの入手が容易になったこともあり，文書処理研究に多くの研究者が参入するようになった．

特許文書検索に関しては，1970年に特許法が改正され，特許情報を計算機により迅速に処理・提供するための機関として（財）日本特許情報センタ（Japatic）が設立された．1978年には特許庁からの情報提供のもと，日本初の特許情報オンライン検索システムがサービスとして展開された．当時の特許出願はまだ紙ベースであったため，検索できる情報は非常に限られているとともに，アクセスできる人も少数に限られていた．しかし，1990年に特許の電子出願が施行されてからは，特許文書を電子データとして蓄積できるようになった．1999年には特許電子図書館IPDL（Industrial Property Digital Library）サービスが開始され，一般の人々でもインターネット経由で特許情報に容易にアクセスできるようになった．

5.2　文書検索の研究動向

5.2.1　文書検索モデル

大量文書の中から検索条件に適合する文書を検索する際に，ユーザによって入力される検索条件は大きく，自然言語文，索引語集合，論理式の三つの表現方法に分けられる．

自然言語文は，頭に思い浮かんだ検索要求を文章として入力して検索条件としたり，既存の文章を流用して検索条件としたりするものである．自然言語文の想起にかかる時間は少なく済むが，その内容を文書検索システムに正しく理解させるためには高度な自然言語処理技術が不可欠である．

索引語集合は，予め検索対象文書群から収集された索引語リストの中から，ユーザの欲しい情報に関連する索引語を選択入力するものである．検索漏れが少ない反面，索引語の選択にノウハウを必要とするために，ユーザにとって最適な索引語を選択することが困難な場合がしばしば起こる．また，選択される索引語の数が少ない場合，適合文書を高精度に特定することができないことが多い．

論理式は，検索で使うタームの関係を論理演算子（AND/OR/NOT）によって規定するものである．ターム間の関係を規定できるという利点を持つ反面，ターム毎の重要度を指定で

[13] NTCIR　http://research.nii.ac.jp/~ntcadm/index-ja.html

きないことや，論理式を組み立てるのにノウハウを必要とするため，検索条件の作成に時間がかかるという欠点がある．

検索条件から得られるタームと，検索対象となる各文書から抽出されたタームの比較によって検索を行うことを前提とした場合の検索モデルとしては，主に以下の 3 種類が普及している[108][42]．

(1) ブーリアンモデル

検索条件として表現された論理式が真となる文書を検索する方式であり，現在最も多くの文書検索システムで採用されている．ここでは，論理式を構成する各タームが文書に存在するか否かをチェックすることにより，論理式の真偽を判別する．本モデルでは，ターム間の関係を規定できる反面，論理式が真か偽かのみによって適合文書を判別するため，検索結果文書を順位付けすることが困難である．

(2) ベクトル空間モデル

検索条件及び検索対象となる文書をそれぞれ重み付きターム集合からなるベクトルで表現し，ベクトル間の類似度（内積や余弦）を算出する方式である[96]．類似度の大小によって検索結果を順位付けることができるとともに，検索条件として使用するタームまたはその重みを修正して再検索することもでき，ユーザの意図を反映させた検索が行える．

ベクトル空間モデルにおいて，文書を重み付きターム集合で表現する手法として，TF-IDF 法が広く採用されている[86]．TF-IDF 法は，一文書内のターム出現頻度（Term Frequency，TF）及び検索対象となる文書群におけるターム出現文書数の逆数（Inverted Document Frequency，IDF）の積によってタームの重みを定義する．すなわち，ある文書において何度も繰り返し記述され，かつ，少ない文書にしか出現しないタームをその文書を特徴付ける重要なタームであると定義している．

ベクトル空間モデルでは，各タームは互いに直交しているという仮定を置いているが，この制約を緩和してターム間の関連を文書間の類似度算出に反映させる方式として，LSI（Latent Semantic Indexing）が提案されている[10]．LSI では特異値分解（Singular Value Decomposition，SVD）によって，ターム空間を縮約する．縮約された空間を構成する各次元はタームの概念に相当すると考えることができる．また，これと似た発想として，タームの代わりにタームの意味属性を用いてベクトルを生成する研究アプローチもある[27]．

(3) 確率モデル

検索条件と検索対象文書が与えられた時に，ある文書が検索条件に適合している確率を推計する方式である[86]．その文書に出現するタームと出現しないタームの両方の情報を用いて適合確率を算出するが，当初は確率変数としてタームが出現する/しないの 2 値を用いていた．その後，確率モデルに TF-IDF 法を組み込んだ文書検索方式が提案された．本方式は英シティ大学で提案されたシステム Okapi のバージョンの一つである BM25 として TREC-4 で報告され[87]，ベクトル空間モデルと同等若しくはそれ以上の精度を実現する方式として学術界で広く採用されている手法となっている．

また，確率モデルの一種として，推論ネットワーク（Bayesian Network）を用いた検索手法も提案されている[110]．推論ネットワークはノードが命題論理の変数または定数，ノード間のリンクが命題間の依存関係を示す有向ネットワークである．Turtle と Croft は，推論ネットワークのノードとして文書，ターム，検索条件を配置し，ある文書が検索結果に含まれる時に検索条件が真となる確信度を算出することにより検索結果を出力する方式を開発し，検索システム INQUERY として実現している[110]．従来の確率モデルでは，ターム間の独立性を仮定していたのに対して，推論ネットワークではターム間の従属性をモデル化できるという特徴がある．

5.2.2 インデクシング方式

初期の文書検索システムにおけるインデクシング方式は，予め用意された統制語の中からその文書を特徴付ける統制語を選択して付与し，検索時には統制語の一覧をユーザに提示し，選択された統制語をインデクスとして含む文書を検索結果として出力する方式であった．本方式によれば，タームの表記揺れがなく，ユーザによるタームの入力誤りも影響しないため，ユーザから指定された条件に合致する文書を漏れなく検索できる．しかし，個々の文書に統制語を付与する作業は文書が膨大になるほどコストがかかることと，ユーザの検索意図に合致する統制語がない場合，検索ができないという欠点がある．

その後，形態素解析をはじめとする文章解析技術の発展により，文書に含まれるターム（非統制語）を自動抽出してインデクスを自動生成する方式が主流になってきた．予め単語辞書を用意し，形態素解析によって文書中の単語を切り出し，不要語除去などの処理を経た後に文書を特徴付けるタームを特定する方式である[41]．本方式によれば，ユーザから指定される任意のタームを含む文書を検索することができ，ユーザの検索意図をより反映した検索が実現できる．その反面，単語の切り出しの精度が形態素解析のアルゴリズム及び単語辞書に登録される単語の数と質に左右されるため，単語辞書を定期的に更新しな

ければならないという欠点がある.

上記辞書保守の課題を克服し，更に形態素解析の誤解析による誤検索を解決する方式として，N-gram 方式が提案されている[92][61][71][78][79]．これは，文書中で連続する N 個の文字を一つのタームとみなし，これをインデクスとして持つ方式である．本方式によれば，検索条件として与えられたターム文字列を漏れなく特定できるため，検索結果に漏れがなくなる．また，単語辞書が不要であるため，その保守コストが不要となる．しかし，タームの意味を無視してタームを認定しているので，例えば動物の「トラ」を検索条件とした場合に，「トラック」「トラフィック」といったノイズタームを含む文書まで検索されてしまうという欠点がある.

現在市販されている文書検索システムは，形態素解析方式または N-gram 方式を採用しているものがほとんどであり，検索対象文書の規模や検索目的，計算機リソースの制約などに応じて使い分けられているのが現状である.

5.2.3 ターム抽出及び重み付け方式

これまでに文書検索に係る様々な方式が提案されているが，これらの方式にほぼ共通しているのが，文書に含まれるタームを利用している点である．また，多くの検索方式ではタームの重みを利用して検索を行っている.

文書からのターム抽出及び重み付けの歴史は古い．Luhn は，高頻度及び低頻度のタームを除いた中頻度のタームが索引語として適切であると主張している[50]．また Maron は，中頻度のタームが索引語として適切であるという仮定のもと，Zipf の法則に基づき，頻度 1 の語の異なり数から中頻度のタームを特定することができると提案している[52].

ターム抽出及び重み付け方式に係る最近のアプローチとしては，特定の言い回し[39][59]や，ターム間の関連[12][63][65][72][81][112][118]または文書間の関連[98]，文書/文章構造[23][95][59][38]，付属語情報[76]，意味属性体系[37]やシソーラス[80]などに着目する研究が多い．木本は，特定の言語表現に着目して新聞記事からキーワードを自動抽出する方式を提案している[39]．中川らは，単名詞のバイグラムから得られる単名詞の統計量，すなわちある名詞が複合名詞を形成するために連接する名詞の頻度を用いて専門分野コーパス文書から専門用語を抽出している[72]．杉山らは，一つの文書だけでなくハイパーリンクで結ばれた文書間の内容を加味した TF-IDF 法の改良を提案している[98]．河合らは，文書構造に着目して箇条書きや表形式文書に書かれた内容を抽出する試みを行っている[38]．また，白木らは，文献の目次に着目して自然言語入力と柔軟な照合によって検索を行うシステムを提案している[95]．松尾らは，文書中の頻出語と各語の共

起頻度の偏りの大きい頻出語を重要なタームとしている[63].

5.2.4 検索アルゴリズム

文書検索システムが検索対象とする文書にはさまざまな種類がある．文書検索精度を向上させるためには，これら文書の特性を考慮する必要があるため，その検索アルゴリズムは文書の種別によって多少異なっている．ここでは検索アルゴリズムの研究動向について，検索対象文書の種別毎に概観する．

(1) Web ページ

Web ページ検索の代表は検索エンジンである[114].代表的な検索エンジンの検索方式を大別すると，ターム検索型，ディレクトリ型，引用分析型に分けられる．

ターム検索型は，検索条件としてターム集合を入力して合致する Web ページを検索するという一般的な検索方式である．最も単純な方式であるが，ページ数が大規模になると検索結果が膨大な数となり，所望のページに辿り着くのが困難になるという欠点がある．

ディレクトリ型は，予め定義しておいた分類体系に沿って個々の Web ページを分類しておき，検索時にそのディレクトリを辿ることにより，検索対象を絞り込むという検索方式である[29]．検索対象となる文書件数を減らすことができ，ターム検索型の欠点を克服できると考えられるが，ディレクトリ体系がユーザの検索意図に合致していない場合，ディレクトリによって検索対象を絞り込むことができないという欠点がある．また，予め各文書をディレクトリに仕分けしておく必要があるため，文書登録時のコストが大きくなる．

引用分析型は，Web ページを互いに独立とみなすのではなく，リンクによって互いに関連付けられているとするもので，多くのページから引用されているページや，多くのページから引用されているページが引用しているページを重要視することにより，検索結果を並べ変える方式である[83][114]．現存の検索エンジンの中では，引用分析型が最も精度が高いとみなされているが，他のページからは引用されないが重要であるようなマイナーなページの検索には有効ではない．そこで，Web ページ間の類似（セマンティクス）に基づく相互評価を行うことによって検索精度を上げる試みもなされている[3].

Web 検索エンジンは精度と性能のみが重要視されがちである．しかし，ある特定の地理的範囲またはある特定の言語に特化して，それらに関連する情報をきめ細かく提供することを重視している検索エンジン[91]や，情報を欲するエンドユーザと情報を提供する組織の両方の要求を満足させるためのサイトナビゲーションを重要視する検索エンジン[100]などもある．更に，検索結果をページ単位で出力するのではなく，検索要求に対する端的

な回答内容のみを提示する，いわゆるQ&Aシステムとしての検索エンジンを目指しているものもある[18]．Q&Aシステムは Web ページのみを対象としたものではないが，情報量の豊富さから Web ページを対象とした研究が多い．

(2) 科学技術文献/新聞記事

論文をはじめとする科学技術文献や新聞記事の検索は評価用コーパスデータの取得が比較的容易であったこともあり盛んに行われてきた．新聞記事に出現するタームの共起頻度[94]や出現位置[94][69]，分野情報[69]に関する情報から，関連する新聞記事を検索する手法が提案されている．大山らは，大規模学術情報データベースを対象とした情報検索システムを提案している[82]．堀井らは概念シソーラスを用いて単語の意味を考慮した，科学技術文献の検索方式を提案している[26]．

(3) 設計文書/議事録/ノウハウ事例

これらの文書は学術的な研究対象になることは比較的少ないが，企業内では設計文書[66][67]やマニュアル，議事録[45]，出張報告や各種のノウハウ[53]などの情報を共有して生産性を向上させるべく，これらの文書の収集及び検索に関する研究が継続的に行われている．設計文書では，同一の作番にある文書間のリンク情報や，作番は違うが開発内容が類似している文書間のリンク情報などを用いた検索方式の研究がなされている[66]．また議事録では，複数回に渡って行われたレビュー結果の変遷を串刺しに検索する機能などが提案されている[45]．更に機械設計ノウハウの検索では，要するにそのノウハウは何を言っているのかを端的に示したフレーズを用いてピンポイントに検索する方式などが提案されている[53]．

(4) アンケート/クレーム/FAQ

アンケートやクレーム情報は顧客から直接得られる情報として重要であり，製品やサービスの質の向上には欠かせない．また，FAQ（Frequently Asked Questions）は企業コールセンタなどへの顧客からの問合せに迅速かつ的確に対処するために有効な情報として重要視されている．他の文書に比べてこれらの文章は短くかつ抽象的であることが多いため，タームの有無だけでなく，タームの共起や主語－述語－目的語の対応関係といった統語的な情報や，付属語情報を加味した文章解析手法[76]が提案されている．乾らは，自由回答アンケートの表層表現から回答者の意図を抽出する方式を提案している[30]．また，FAQ 検索では，顧客対応時の時間的制約の問題も絡み，如何に所望の FAQ を探し出して顧客に回答するかが重要であるため，単なる検索方式だけでなく検索結果を絞り込む方式についての研究もなされている[65]．

(5) 特許文書

特許文書については5.4節で詳説する.

(6) 人の情報

上記のような文書情報ではなく，人の情報を検索するものである．具体的には，ある知識に詳しいエキスパートを検索したり[97][28]，スケジュールの空いている人を検索したりする．稲子らは，検索で用いるターム集合と同一の空間に著者情報を配置することで検索条件に最も合致する人を検索する方式を提案している[28].

文書種別に依存しない文書検索の研究アプローチとして，文書検索結果から得られる情報をもとに検索をやり直すことによって検索精度を向上させる適合性フィードバック方式が盛んに研究されている[109][43][74][75][73]．また，入力文章内容からその分野を推定し，検索対象を絞り込んで検索する方式[48]が提案されている．更に，検索条件の作成を支援する機能についても，検索結果文書集合に含まれるタームの共起性からターム間の関係をネットワークとして可視化するなどの研究成果がある[77][84].

5.3 分類自動付与の研究動向

分類自動付与は類似文書検索と技術的に共通している部分が多い．すなわち，分類付与対象となる入力文書からその特徴を表すタームを抽出したり，重み付けをしたり，文書内容に合致する分類を特定したりする部分である.

分類付与におけるタームの重要度判定の考え方として，あるタームの分類別出現頻度に着目し，それが一様であるならば一般語，偏っているならばその分類を特徴付ける重要語であるとみなす方式が広く知られている[70]．また，ターム抽出において，分類を特徴付けない一般語は不要語としてターム集合から除外している方式が多いが，一方で Riloff は，このような不要語の除去が分類付与精度に悪影響を及ぼす場合があることを実験で検証している[85]．不要語除去の妥当性についてはケースバイケースであることが多く，その是非については議論が収束していないのが現状である.

分類付与において，その分野を特徴付けるタームの多義を解消するとともにその重要度を的確に判定するための手掛かりとして，ターム出現頻度[1][47]のほか，共起情報[12][118]や意味属性[37][33]，格構造[6]，同義語[111]，シソーラス[21]，語義文章[19]，文書構造や文章構造[116][23]，特許文章に特有の言い回し[116]，文書クラスタリング[104]，重み付け学習[20]などに着目した取り組みがある．湯浅らは，名詞の出現頻度及

び共起関係から分類知識を自動生成している[118]．藤井らは，同一段落内の単語の出現状況に応じて同一タームをあたかも別のタームであるかのように扱うことにより多義解消する方式を提案している[12]．一方，サポートベクタマシン（SVM）やブースティングといった機械学習方式を分類付与に導入した研究が 1990 年代後半から非常に盛んになってきている[7][46][101][102]．

重み付きターム集合による分類付与以外の方式としては，Hayes らがルールベースで新聞記事に分類を付与するシステムを実用化している[24]．また，桂田らは分類誤りに対してルールを自動修正する方式を提案している[36]．しかし，ルールの作成・修正には専門家の介在を前提としている．また，特許のように技術分野が広範囲である分類に適用するにはルールが複雑化するために十分な精度が出ないか，精度が出たとしてもそれらのルールだけで分類を自動付与できる文書はかなり少ないと思われる．したがって，ルールベースの付与方式はターム集合を用いた方式と併用することが望ましいと考える．

1.4.1 節で比較したように，分類自動付与方式は大きく KNN 法に代表されるように類似度の高い文書を算出して，それに付与されている分類を出力する方式[4]と，文書と分類の間の類似度を算出する方式がある．KNN 法は，類似文書検索システムと同一のインデクスを用いて容易に実現できること，また一つの文書に複数の分類を付与する場合の再現率が比較的良いことなどから一般的に用いられる方式であり，NTCIR-5 特許分類サブタスクに参加した多くのチームが KNN 法を採用している[32]．一方，文書と分類の間の類似度を算出する方式は，分類を特徴付けるターム集合が分類毎に定義されるので，分類付与精度を更に向上させるためのデータチューニングが比較的容易である，多くの文書からタームを抽出するので表記揺れによるターム照合漏れの影響が小さい，その分類を付与すべきと出力した根拠を比較的提示しやすい，といった長所がある．

5.4 特許文書処理の研究動向

特許文書処理の研究が学術的に本格的に採り上げられるようになったのは,2000 年前後である．1.1 節で述べた社会的背景や，1990 年に施行された特許の電子出願によって電子文書が大量に蓄積されてきたことなどが研究活発化の契機として挙げられる．また，SIGIR2000[14]や ACL2003[15]など，海外の著名な学会が主催する特許文書処理に関するワークショップが開催されたことや，NTCIR-3 で特許検索タスクが立ち上がり，10 年分の特許文

[14] SIGIR2000 特許検索ワークショップ http://research.nii.ac.jp/ntcir/sigir2000ws/

[15] ACL2003 特許コーパス処理ワークショップ http://www.slis.tsukuba.ac.jp/~fujii/acl2003ws.html

書データ及び評価データセットが整備・提供されるようになって研究環境が整ってきたことも研究の活性化に大きく貢献している．

特許文書検索が他のテキスト文書検索と比べて異なる点としては以下が挙げられる．

(1) 文書内容を特定するタグが存在しており，書くべき内容が決まっている．
(2) 文章の長さが比較的長い．ただし，文章の長さは文書によってばらばらである．
(3) 不特定多数の著者によって執筆される．
(4) 特許文書の中核をなす請求項の記載において固有の構文的特徴が見られる．

ここ数年，特許文書処理に係る研究テーマとして広く採り上げられているのが，(1)類似特許文書検索の精度向上技術，(2)特許分類の自動付与技術，(3)特許分析支援技術である．これらはどれもNTCIR特許検索タスクでも採り上げられた重要なテーマである．これらのどの研究を行うにせよ，上記四つの特許文書の特徴を考慮すべきである．以下では，上記研究テーマのうち，類似特許文書検索及び特許分類自動付与に係る研究動向について詳しく述べるとともに，本研究のアプローチと比較する．

5.4.1 類似特許文書検索

ここでは，自然言語文章を入力として，その文章に関連の深い特許文書を検索する精度を向上させる研究の動向について述べる．

NTCIR-4[34]及びNTCIR-5[35]における特許検索タスク[13][14]は,請求項文章を入力としてその発明内容を無効化する特許文書を検索する精度を競うものである．精度向上のために各参加チームが採用した主なアプローチとしては，(1)クエリ拡張[15][44]，(2)分類情報の利用[55]，(3)特許文書の構成[55][31]や入力請求項文章の構造または記述特性の利用[55][44][106]，(4)適合性フィードバックの適用[22]などがある．

本研究では，第3章で詳述したように上記(3)に着目し，(a)検索の漏れ防止及びノイズ防止の両方に着目した段階的検索方式，(b)請求項の記述特性を考慮した，出現頻度を利用しないターム重み付け方式，(c)尺度表現に着目したターム重み付け方式を特徴とする類似特許文書検索方式を提案している．以下，これら3点に焦点を当て，これらに関連する研究動向及び本研究アプローチとの違いについて述べる．

(a) 段階的検索方式（本研究の方式の詳細は3.2節参照）

Clarkeらは,第一段階で検索された検索結果候補を第二段階において並べ替える指標として，タームの近接性と共起性に基づく“cover density ranking”という指標を提案し

ている[9]．またBearらは，第一段階での検索結果のポストフィルタリングにおいて，情報抽出のスキームを導入している[5]．Sumnerらは，検索結果を絞り込むことによって適合性フィードバックの効果が上がると報告している[99]．これらに対して本研究では，請求項の文章構造に着目してタームの抽出範囲を限定するとともに，特許文書構成に着目して特許文書の検索範囲を限定することによって得られる第二段階検索結果スコアを，第一段階で得られる検索結果スコアにマージする方式を採用している点が異なる．

水野は，三段階検索方式を導入している[64]．第一段階では，請求項において発明の前提を表す記述部分（「前提部分」，1.3.2節参照）を使って，特許文書の構成要素のうちの請求項及び要約を検索対象として検索を実行する．第二段階では第一段階での検索結果上位M件のみを対象として，要約の中で目的を表す記述部分を使って，要約を検索対象として検索を実行する．第三段階では第二段階での検索結果上位N件を対象として，請求項において発明の特徴を表す記述部分（「特徴部分」，1.3.2節参照）を使って，特許全文を検索対象として検索を実行する．そして，最終的には第三段階で得られた検索結果が最終の検索結果となる．これに対して本研究では，最後の検索段階における検索結果を最終の検索結果とするのではなく，第一段階での検索スコアをベースとして，第二段階での検索スコアをこれに上乗せするという形で最終の検索スコアを算出し，そのスコアをソートすることによって最終の検索結果を得ている点が異なる．

伊藤は，特許文書全文を検索対象とした時に得られる検索結果と，請求項及び要約のみを検索対象とした時に得られる検索結果をマージすることによって最終の検索結果を算出している[31]．しかし，ここでは入力文章からのターム抽出及び重み付け手法は同一のアルゴリズムを使用しているのに対して，本研究では入力文章からのターム抽出範囲を段階毎に変えているという点が異なっている．

小西らは，入力請求項文章を発明構成要素毎に分割し，分割された個々のフレーズ毎に検索を実行し，それらの検索結果をすべてマージすることによって最終の検索結果を得る方式を導入している[44]．ここでは，入力文章の単位は変えているが検索対象範囲はどれも共通であり，検索段階毎に入力文章の単位及び検索対象範囲の両方を変えている本研究の方式とは異なっている．しかし，請求項を発明構成要素の集合として捉え，個々の検索結果から最終的な検索結果を導く手法は，特許検索の専門家の作業手順に類似していることから，このような検索アプローチは重要視すべきと考える．

請求項を個々の発明構成要素に自動分割する手法としては，新森らの研究がある[93]．ここでは，典型的な記述パターンに着目して発明構成要素の境界を特定し，それらを組み合わせて請求項の構造を体系化する試みがなされている．また，高木らも請求項の発明構

成要素毎に検索を行う方式として，各構成要素の重要度を算出し，それを重みとして検索結果の積和を求める方式を提案している[103]．しかし，一般に構成要素に限定した検索はノイズが多く含まれるため，本研究で提案するように予め検索対象を絞るステップが最初に必要であると考える．

(b) 出現頻度を利用しないターム重み付け方式（本研究の方式の詳細は 3.3.1 節参照）

Sarasua らは多言語情報検索において，技術分野を単位とした出現文書数(IDF)とタームの出現位置を用いて特許文書全文から抽出されたタームへの重み付けを行っている[90]．これに対して本研究では，請求項の記述特性，すなわち，文章が短い，指示語が使われず同じ言葉が冗長に繰り返される，発明対象を表すタームの出現頻度が少なくなる傾向にある，という三つの特性を踏まえた上で，請求項から抽出されるタームへの重み付けは出現頻度を利用しない方が良いと主張している．

竹内らは，請求項におけるタームの出現位置に着目し，請求項の中央近辺に出現するタームが重要なタームであると仮定している[106]．この考え方は，出現頻度を利用すべきではないという本研究と似た着眼であり，本研究の方式と組み合わせることにより，よりターム重み付け精度が向上する可能性がある．

(c) 尺度表現を用いたターム重み付け方式（本研究の方式の詳細は 3.3.2 節参照）

多くの質問応答システムや一部の文章要約システムでは，文章から 5W1H 情報を抽出するために，タームの意味属性を用いている．本方式でも，あるタームが「速度」や「膨張率」などの尺度を表すタームであるかを判定するために，タームの意味属性に基づいて尺度表現語を定義している．ただし本研究では，請求項において発明の特徴を表すためにどんな種類のタームが使われるかという傾向分析結果から尺度表現が重要であるという仮説を導いており，この手法は特許文書からのターム抽出に特化した方式であると考える．

ところで，類似特許文書検索では他の文書検索にはない特徴が一つある．それは第 4 章で詳述したように，ある出願特許文書に書かれた発明を無効化する特許文書を検索する際に，その出願特許文書と無効化特許文書とが同一の発明者または出願人である場合が実に約 21%であるという事実である．計量文体学において，「同一の執筆者が執筆する文章は使用する語彙や構文，文章構成が類似することが多い」という知見がある．この知見と上記事実を踏まえると，「出願特許文書とその無効化特許文書が同一の発明者または出願人である場合とそうでない場合で，類似特許文書検索の精度的振る舞いが大きく変わる可能性がある」という仮説を得ることができる．本研究では第 4 章において，この仮説が正しいことを実験によって立証した．

文章の定量的な言語的特性から執筆者や文章の種別を類推する研究は計量文体学では盛んに行われている．すなわち，言語学的特徴として，文や単語の長さ[2][8][25][68][115]，タームの頻度[2][25][68] や品詞（名詞や接続詞）[8][115]，文字種[8][115]，構文[17][68]，文中の出現位置[68] などに着目して，文章の執筆者を推定したり，文章のタイプ(論文，記事，Web ページ)を識別したりする研究である[11][49]．本研究ではこれらと同様のアプローチでタームの使用傾向を分析しているが，執筆者のレベルだけでなく，出願人という組織のレベルで分析している点が異なる．

5.4.2 特許分類自動付与

吉田は，特許分類自動付与において二段階付与方式を提案している[117]．第一段階では，IPC の上位分類であるメインクラス（118 分類）毎にその分類を特徴付けるタームを保持し，付与対象特許文書中のタームとの類似度を算出することにより，類似度の高い上位 3 分類を特定する．第二段階では，この上位 3 分類が付与された特許文書のみを対象として，KNN 法によって類似度の高い上位 50 件の特許文書を検索し，それらに付与されたテーマ分類毎に類似度を積算し，積算値の高いテーマ分類を出力する．また，分類アルゴリズムの改良方式として，国内優先権主張，分割出願，出願人 IPC といった，特許固有の属性情報を加味した方式を模索している．しかし，この方式ではタームの重み付けにおいて出現頻度を用いており，本研究のようにタームの出現位置や出現共起を用いることについて言及されていない．

Larkey は入力文章に類似する特許文書を検索する際に，入力文章内容から技術分野を KNN 法により推定することにより検索対象を絞り込む方式を提案している[48]．特許文書を構成するセクション毎に重みを設け，あるタームがそのセクションに出現する頻度にその重みを掛け合わせた値をそのセクションにおけるそのタームの重要度とし，すべてのセクションについてその重要度の合計値をその文書におけるそのタームの重要度としている．タームの出現位置を考慮して重み付けをしている点では本研究と共通しているが，セクション間のタームの出現共起については考慮されていない点で本研究と異なる．

余田らは，TF-IDF 法で算出されるターム重みを，特許文書の構造的・構文的特徴から算出されるターム重みで補正する方式と，タームの出現共起情報を用いてターム重みを補正する方式を提案している[116]．タームの重み付けに関する考え方は本研究と共通しているが，ターム重み付けのベースとしてここでは TF-IDF 法を使っているのに対して，本研究では TF の代わりに出現位置情報及び出現共起情報のみを用いている点と，本研究では IDF の代わりに分類別のターム重みの偏りを用いて類似度を算出している点が異なる．

第６章　結論

本章では，本研究の成果及び今後の課題について総括する．

6.1　研究の成果

本研究では，特許文書を対象とした分類自動付与技術及び類似特許文書検索技術について検討した．

第 1 章ではまず，特許をはじめとする知的財産権の保護・活用に係る社会的背景及び特許文書を含む文書検索に係る技術的背景について概説した．まず，特許庁における特許審査業務及び企業における特許戦略立案業務において，特許文書検索作業が大きなウェイトを占めており，迅速で的確な検索作業を実現するシステムが要望されていることを述べた．

そこで，本研究では研究目的として，(1)特許分類自動付与技術の確立と，(2)類似特許文書検索技術の確立，の 2 点を挙げ，これらの精度を向上させる方式について検討することとした．

次に，特許分類自動付与研究に係る基本課題として以下の 4 点を挙げ，それぞれに対して以下の解決策を採ることとした．

［課題 1］特許文書に記載される発明内容をどのように記述するか？
［解決策］重み付きタームの集合として発明内容を記述する．

［課題 2］特許文書に記載される発明内容をどのように特定するか？
［解決策］特許文書の構成と記載文章の構造（構文的特徴）に着目して発明内容を端的に表すタームを抽出する．また，発明が適用される対象物または技術分野について端的に記載した箇所を重要視する．

［課題 3］各分類が網羅する技術分野範囲をどのように定式化するか？
［解決策］過去に分類済みの大量の特許文書からその分類固有のターム特性を抽出する．また，分類定義文章から分類に関係の深いタームを抽出して補完する．

［課題 4］発明内容と技術分野の類似性をどのように判定するか？
［解決策］付与対象文書のタームと各分類を特徴付けるタームの間の類似度を算出する．

上記解決策を決定する際には，実用化の観点から以下の 3 点を具備要件として挙げ，こ

れらを満たす解決策を採用した.

［具備要件 1］上位分類の自動付与では，分類をノイズなく特定できること（正確性）
［具備要件 2］詳細分類の自動付与では，付与結果の根拠を提示できること（透明性）
［具備要件 3］分類付与に係るデータやプログラムがカスタマイズ可能なこと（拡張性）

一方，類似特許文書検索に関して以下の 3 点の基本課題を挙げ，それぞれ以下の解決策を採ることとした.

［課題 1］特許文書に記載される発明内容をどのように記述するか？
［解決策］重み付きタームの集合として発明内容を記述する.

［課題 2］特許文書に記載される発明内容をどのように特定するか？
［解決策］特許文書の構成と請求項の文章構造に着目してタームを抽出し，重みを付与する. 発明の技術分野に係るタームよりも，発明の特徴を表すタームを重視する.

［課題 3］入力となる発明内容と特許文書の間の類似度をどのように算出するか？
［解決策］重み付きターム集合間の類似度を算出する．また，類似する特許文書を段階的に絞り込む．更に，異なる観点で算出された類似度をマージする.

第 2 章では，特許文書への分類自動付与技術について述べた．特許分類は特許文書検索における検索対象の絞り込みとして非常に有効な手掛かりとなる．ここでは，特許分類体系の上位分類であるテーマ（2,815 分類）及び審査室（38 分類）を付与対象分類とし，特許文書の構成及び請求項構造に着目した以下の 4 点を特徴とする分類自動付与方式を提案した.

(1)特定の特許文書タグのみから発明の技術分野を特定するタームを抽出する方式
(2)タームの出現位置及び出現共起に基づいてタームに重みを付与する方式
(3)大量の分類付与済み特許文書と，各分類の適用範囲を文章で規定した分類付与マニュアルからそれぞれ抽出したタームを統合し，各分類を特徴付ける重み付きターム集合からなる分類知識を自動生成する方式
(4)タームの分類別出現傾向及び分類体系の階層性に着目した類似度算出方式

新規特許文書にテーマ及び審査室を自動付与する精度評価実験を行った．その結果，分類を 1 種類ずつ付与した場合の正解率としてそれぞれ 61.6%, 83.3%, 分類を 3 種類ずつ付与した場合の正解率としてそれぞれ 82.8%, 96.0%を得た．また，分類付与済み特許文書データに加えて分類付与マニュアル文章を教師文書データとして利用することにより，正解率を 2.2%から 10.5%までの範囲で向上させることができ，分類付与マニュアル文章を利用

する有効性を確認した．更に，分類知識保守の観点から，分類知識生成に必要十分な教師文書データ量を検証した結果，1 分類あたり約 1,000 件の分類付与済み文書データが必要であることを確認した．更に，審査室を対象として全自動分類付与システムとしての実現可能性を検証した結果,審査室が 1 種類だけ付与された評価データ(64.5%)のうちの 94.2%について，自動付与結果として出力された上位 3 個の審査室の中に正解審査室を出力できること，評価データ全体の 80.0%については，自動付与結果上位 3 個の審査室の中にすべての正解審査室を出力させることができることを確認した．

第 3 章では，代表的な特許請求項である請求項 1 の文章を入力として，その発明内容を無効化する特許文書を検索する類似特許文書検索の精度を向上させる方式を提案した．検索漏れ防止及び検索ノイズの低減の両方に対処すべく，特許文書の構成及び請求項文章の構造に着目した検索方式として，(1)検索段階に応じて検索タームの抽出範囲及び検索対象を変え，各検索段階における検索結果から最終的な類似度を算出する二段階検索方式，(2)請求項文章の記載に係る言語的特性を踏まえた，出現頻度を利用しないクエリターム重み付け方式（TF1），(3)発明の特徴を表す尺度表現に着目したクエリターム重み付け方式の 3 方式を提案した．

これらの方式の有効性を検証すべく，約 170 万件の公開特許公報データに対して，NTCIR-4 特許検索タスクのフォーマルランの課題データ 103 件などを使用した評価実験を行った．その結果，評価データセットによって傾向にばらつきがあるものの，出現頻度を利用しないクエリターム重み付け方式及び二段階検索方式は，全体として検索精度（平均精度）を向上させる効果が高いことを確認した．また，尺度表現に着目したクエリターム重み付け方式は，平均精度の改善には貢献しないものの，正解特許文書の検索順位を全体的に押し上げるのに非常に有効な方式であることを確認した．

第 4 章では，ある出願特許の発明内容を無効化する特許文書が，出願特許と同一の発明者または出願人によるものである場合が多いという事実を重要視し，出願人の同一性が類似特許文書検索精度の評価に大きく影響することを実験により示した．

まず，クエリとその無効化特許の出願人に関する傾向について，(1)文書属性，(2)使用タームの共通性，(3)検索の難易度という 3 つの観点から定量的に分析した．その結果，(a)クエリとそれを無効化する特許文書の出願人が同じであるという現象が無効化特許文書件数の約 21%で起こっている，(b)この現象は技術分野や特定の出願人に偏って起こっている現象ではない，(c)出願人が違う無効化特許文書の検索は，出願人が同じ無効化特許文書の検索に比べて難しい，(d)これまでに提案されてきた検索方式の精度的振る舞いは，出願人の同一性によって異なる場合がある，という知見を得た．

次に，これらの知見を適用した，類似特許文書検索方式の組み合わせ手法を提案した．すなわち，出願人の同一性の観点から個々の検索方式の有効性を評価し，その結果を踏まえて複数の検索方式を組み合わせて最適な検索方式を得る手法を提案した．NTCIR-5 特許検索タスクの評価データを用いて本手法を評価した結果，出願人の同一性を考慮しないで検索方式を評価する場合と考慮した場合で，検索精度が最良となる検索方式の組み合わせが大きく異なることが分かった．このことから，類似特許検索精度の有効性を正確に評価するためには，出願人の同一性を考慮すべきであるという結論を得た．

第 5 章では，分類自動付与技術及び文書検索技術を中心にこれまでの研究動向について概観するとともに，本研究との比較を行なった．分類自動付与及び検索精度の向上技術については，Web をはじめとして種々の文書を対象とした種々の方式が提案されているが，技術的にはまだ発展途上段階にあると言える．

6.2　今後の課題

6.2.1　分類自動付与技術に係る今後の課題

分類自動付与精度を更に向上させるために，今後検討すべき技術課題としては以下が挙げられる．

(1) ターム抽出・重み付け精度の向上

分類自動付与精度向上の最大の課題は，ターム照合漏れの防止とターム重み付けの最適化である．ターム照合漏れの最大の原因は，特許文書の執筆者が不特定多数であるため，同じ概念に対して各執筆者が使う言語表現が異なることにある．したがって，ターム照合漏れを防止する方策としてはシソーラスの利用が有効と考えられる．しかし，広い技術分野を網羅する特許分類体系との親和性の高いシソーラスを構築・保守していくのは，コストが高くあまり現実的でない．代替案として，教師文書データから互いに意味的関連の深いターム対を自動抽出して，付与対象の特許文書から抽出されるタームを拡張・補完することにより，ターム照合漏れを軽減することができると考える．これは一種のタームの多義解消である．ただし，一般にこのようなターム拡張方式は再現率を向上するための方式であり，本方式で前提としている分類付与担当者への自動振り分けのような適合率重視の分類自動付与には貢献しない可能性がある．

一方，ターム重み付けの最適化の方策としては，請求項の構造をより細かく解析して，タームの重み付けをより緻密に行うことが考えられる．例えば，その発明を構成する要素

は何か，またはその発明を実現する処理ステップは何かを記述している部分を抽出し，そこに含まれるタームの重みを高くするといった方式である．本研究では，発明の対象物または技術分野を端的に表すタームを重要視したが，実際にはこれらのタームだけでは付与すべき分類を一意に特定できない分類もある．どこまで細かく文章構造を解析すべきかは，分類体系の粒度にも大きく左右されるため，ターゲットとする分類体系の粒度を考慮してターム抽出及び重み付けを最適化する必要があろう．

また，ターム間の共起関係を考慮することも重要である．例えば，「画像」や「検索」といったタームはいろいろな技術分野で使われるが，「画像を検索する」というように，二つのタームが動詞＋目的語という関係で記述されている場合，技術分野を大幅に絞り込むことができると考えられる．ただし，このような共起関係を使用する場合，その出現頻度は比較的低くなるため，ノイズの影響を受けやすくなる．したがって，シソーラスなどを利用して表記揺れを極力統一する仕掛けを併用することが不可欠である．

(2) 分類体系の特性を加味した分類付与

分類自動付与精度は分類体系の定義の良し悪しに大きく影響する．各分類が網羅する技術範囲が互いに排他的であり，分類間のタームの共通性が低いほど，計算機による分類自動付与は容易となる．また，各分類固有の特性も影響する．例えば，ある分類をある特許文書に付与すべきかを判定する際に，発明の名称だけを見れば分かるような，自動付与が比較的容易な分類もあれば，請求項の意味を正確に理解したり，本文の詳細まで読解したりしないと判定できないような，高度に知的な判断が必要とされる分類もある．したがって，付与されるべき分類が網羅する技術分野がどのような特性を持っているかを考慮した分類付与方式が必要となる．

また，上記(1)でも述べたが，同じ階層の分類でもその粒度が全く異なる場合がある．特許分類体系においては，ある分類の特許出願件数が増加した場合に，その分類を細分化することがしばしばある．その結果，出願件数の多い技術分野の分類の粒度が他の技術分野に比べて細かくなる傾向がある．例えばテーマの場合，「望遠鏡」に関する発明に対応するテーマは一意に決まる（2H039）が，「カメラ」に関する発明に対応するテーマは，「絞り（2H080)」や「シャッター（2H081)」など数十に跨っている．そこで，このような分類の粒度の違いを加味し，ある共通の発明対象物（上例では「カメラ」）に関する分類を一つの仮想的な分類として捉えて分類を段階的に付与するというような方式も有効であると考える．すなわち，まずこの仮想分類を単位とした分類付与を行い，次に付与された仮想分類の中のどの分類を付与すべきかを判定する方式である．この際，仮想分類を単位とした分類付与では重要であったターム（上例では「カメラ」）が，下位分類の付与ではど

の下位分類にも共通して出現するタームとなり重要でなくなるため，別の観点で重要なタームを特定することが必要となる．

一方，テーマのように分類体系が大規模である場合，自動付与精度が 100%になることは現状の解析技術ではまず不可能であるので，分類付与に正確さを求める場合には，以下に挙げるように分類自動付与結果の正誤をユーザが簡単にチェック・修正できるための支援機能が必要となると考える．

(a) 人手でチェックすべき特許文書の自動選定

分類自動付与結果を人手でチェックする場合，チェックすべき特許文書件数を極力減らすことにより，チェックを効率化できる．すなわち，付与すべきであることが明白である分類については，人手のチェックを介さずに計算機による自動付与結果をそのまま採用し，逆に付与すべきかに関する確信の度合いが低い分類については人手のチェックを受ける．具体的には，分類自動付与の類似度にしきい値を設け，そのしきい値よりも高い値を持つ分類については人手のチェックを受けないで済むようにするという方法が考えられる．また，「発明の名称に『望遠鏡』というタームが出現したら，必ずテーマ 2H039 を付与する」というような信頼性の高いルールを自動抽出し，そのルールを満たす分類については人手のチェックを受けないで済むようにするという方法も考えられる．

(b) 分類自動付与根拠の提示

自動付与結果を人手でチェックする場合，なぜその分類が付与されたかに関する根拠が知りたくなる．分類付与担当者はその根拠を理解して初めて自動付与結果が正しいか否かを判定できる．したがって，自動付与結果を出力する際に，その分類を付与するのに貢献したタームの一覧やそれらの本文中での出現箇所などの根拠情報を提示することにより，本文中のどのあたりの記載部分に着目して分類が自動付与されたのかを，担当者が理解できるようになる．

(c) 誤付与時における正解分類候補の提示

自動付与結果が誤っていると人手によって判定された場合，正しい分類に修正する作業が不可欠となる．その場合，大規模の分類体系の中から適切な分類を探すのは労力のかかる作業であるので，正しい分類の候補をシステム側から提示する機能が必要となる．具体的には，自動付与結果を類似度別に順位付けて出力する機能のほか，自動付与に用いたターム及びその重みを担当者がチューニングして自動付与を再実行する機能，ある分類と意味的に関連の深い分類（例えば分類知識に蓄積されているターム集合の共通性が高い分類）を提示する機能などが有効であると考える．

6.2.2 類似特許文書検索技術に係る今後の課題

類似特許文書検索精度のさらなる改善を実現するためには，何よりもまず，評価環境を整備することが不可欠である．今回の評価実験の結果，評価データセットによって精度の改善傾向にばらつきが見られた．これが何に起因するのかを特定するために，実験結果をミクロに分析する必要がある．その一つとして本研究では，出願人の同一性の観点から分析をしたが，この他に，評価データの規模，評価データの網羅する技術分野の偏り，正解特許文書の定義方法，個々の特許文書の記載の特殊性，検索方式の性質などの観点からも実験結果をミクロに分析することで，検索精度のさらなる改善に向けたアプローチの策定及び正確な有効性の評価に結びつける必要がある．

本研究で提案・評価した類似特許文書検索の精度改善方式を更に高度化するアプローチとしては，以下が考えられる．

(1) 請求項構成要素のさらなる活用

本研究で提案した二段階検索方式では，請求項を前提部分と特徴部分の二つに分け，各段階に応じてターム抽出範囲を変えた．しかし，実際の特許審査における無効化特許調査では，請求項を更に細かい構成要素，すなわち発明の構成手段や処理ステップの単位まで分割し，この単位毎に類似特許文書を検索して，その結果を組み合わせることが多い．そこで，入力となる請求項文章を解析して構成手段や処理ステップ毎に自動分割し，構成単位毎の検索結果を統合することにより，無効化特許の検索精度を改善することが可能となると考える．請求項文章の構成単位の特定は，請求項固有の言い回し（「A 手段と，B 手段と，・・」「A し，B し，・・」など）に着目することによって，ある程度の解析精度が期待できる．一つの請求項文章全体を用いて検索するよりも，その請求項を構成するある特定の構成単位のみを用いて検索した方が，検索精度が改善される場合が多いことが分かっているが，検索精度が改善される構成単位をどのようにして特定するかが難しい．かといって，構成単位毎の検索結果を単純に統合すると，検索ノイズの影響を受けやすい．したがって，請求項から得られる構成要素の情報をどのように検索に活用すべきかについて更に検討する必要がある．

(2) 請求項以外の文章の活用

本研究で提案した類似特許文書検索方式は，クエリとして発明内容を端的に表す請求項 1 の文章のみを用いた．しかし，実際の検索作業では，クエリとして請求項文章だけしか使えないという状況は比較的少なく，特許文書全体を入力として使えることも多い．その場合，請求項以外の文章情報を用いることで，発明内容をより的確に定式化できると考え

る．例えば，請求項文章が非常に短く，検索に用いるタームが十分に抽出できなかった場合には，他の文章部分からタームを補完することができる．この時，請求項の内容に対応する本文中の詳細記載部分を特定し，その記載部分からタームを補完することにより，ノイズタームの混入を防ぐことができる．また，タームの重み付けにおいても，特許文書全体における出現頻度や出現箇所，出現共起の情報を加味することにより，本研究で提案した「出現頻度を用いないターム重み付け方式」よりも適切なターム重み付けが可能であると考える．

(3) 分類・出願人の利用

特許文書の文章だけでなく，特許文書の属性情報として分類や発明者・出願人などの情報を活用することにより，検索精度の向上が期待できる．公開された特許文書には分類が付与されているので，これを検索対象の絞り込みに活用することにより，ノイズ文書を大幅に減らすことが可能である．一方，分類情報が未知の場合には，第 2 章で述べた方式や KNN 法などを用いて付与された分類を活用することで検索結果を絞り込むことができる．

しかし，クエリの発明内容を無効化する特許文書として，クエリと全く異なった分類が付与された特許文書が用いられることがしばしばある．この場合，分類によって検索対象を絞り込むことによって検索漏れが発生する恐れがある．そこで，検索結果文書のうち，クエリと共通の分類を持つ特許文書について，そのスコアを上げることにより，検索漏れを最小限にしつつ，全体の検索精度を向上させることができると考える．また，ある分類に関連の深い分類として，分類を特徴付けるタームが多く共通している分類や，ある分類が付与された特許に対する無効化特許に付与されている分類を補完することも有効である．ただし分類を用いる場合には，どの種類の分類（IPC/FI 体系，テーマ/F ターム体系）を用いるのか，またどの階層の分類を用いるのかについて考慮する必要がある．

(4) 技術分野の特性の利用

本研究で提案した類似特許文書検索方式では，クエリの属する発明技術分野による精度的振る舞いの違いについて考慮していない．しかし，実際には技術分野によって，発明内容を表す重要タームの抽出基準が異なるのではないかと考えている．例えば，化学分野の請求項では，ある物質を構成する物質の組み合わせや含有率が重要であるが，情報処理分野の請求項ではこのような観点では記載されない．また，本研究で提案した類似特許文書検索方式では，クエリ中のタームの重要度を測る手掛かりとして，出現頻度を用いない方式を採用しているが，実際には出現頻度が有効な技術分野と有効でない技術分野が存在すると考えられる．このような技術分野別の特性を更に分析し，クエリの技術分野に応じて検索アルゴリズムを動的に選択するような工夫が必要であろう．

6.2.3　本研究成果の特許文書以外への適用可能性

最後に，本研究で提案した特許分類自動付与方式及び類似特許文書検索方式が，特許文書以外の文書にどの程度適用可能かについて述べる．文書の構成及び文章の構造は，文書の種類によって大きく異なるため，その検索精度を向上させるための方式も大きく異なる．ここでは，特許文書のように，文書構成または文章構造にある程度の統一性が見られる科学技術論文と新聞記事を例に考察する．

科学技術論文は，論文全体の内容を短くまとめた要旨（抄録）と本文から構成される．このうち，要旨では，主に，研究の背景，技術課題，その課題を解決するアプローチ（方式)，そのアプローチの評価結果について簡潔にまとめられている．一方，一般に過去の論文を検索する場合，利用者が今興味を持っている技術課題または解決アプローチと類似する論文を検索したいというニーズは高いと思われる．

本研究で提案した二段階検索方式を科学技術論文に適用してみる．ここでは，要旨をクエリとし，その研究内容，特に技術課題が類似する論文を検索すると仮定する．第一段階では，要旨文章全体を用いて，技術分野を特定するレベルの検索（広く浅い検索）を実行する．第二段階では，要旨文章のうち，技術課題について書かれた文のみをクエリとすることで，技術課題の類似性をピンポイントに比較する検索（深く狭い検索）を行い，第一段階の検索結果を並べ替える．ここで，技術課題について書かれた文は，文頭における接続詞の使い方（逆接の接続詞が使われているなど）や，文末の付属表現（「～してしまう」「～できない」など)，課題について述べるときにしばしば使われるネガティブな表現(「課題」「欠点」「困難」）に着目することである程度特定できると思われる．以上の方法によって，二段階検索方式を科学技術論文の類似文書検索に適用可能であると考える．

二段階検索方式は，新聞記事についても同様に適用できると考える．ここでは，新聞によく見られる 500 字程度の記事文章をクエリとし，そのトピックに類似する過去の記事を検索すると仮定する．類似する新聞記事の検索では，記事の 5W1H（いつ・どこで・誰が・何を・どのように・なぜ）が同じものが重視されると考える．新聞記事では，そのほとんどが第一文に集約して記述され，第二文以降でその詳細について述べている場合が多い．そこで第一段階では，記事文章全体を用いて，記事のトピック分野（新聞の面に相当）を特定するレベルの検索（広く浅い検索）を実行する．第二段階では，記事文章の第一文のみをクエリとすることで，5W1H の類似性をピンポイントに比較する検索（深く狭い検索）を行い，第一段階の検索結果を並べ替える．

科学技術論文及び新聞記事に係る上記検索方式は，これらの文書の分類自動付与にも適用できる．すなわち，上記検索方式で，第二段階で使われるタームの重みを上げることに

よって，分類を特徴付けるターム及びその重みをより正確に記述することができ，検索精度の向上につながると考える．

本研究で提案した類似文書検索及び分類自動付与方式は，特許や科学技術論文，新聞記事のように，文書構成及び文章構造にある程度の統一性が見られる文書に適用可能である．しかし，例えばWebページのような，多種多様な文書構成・文書構造を持つ文書が混在している文書への適用は難しい．文書の種類毎にターム抽出及び重み付け方法を規定しなければならなくなることと，異なる種類の文書を一括して検索する際に，その重みの価値をいかに統一するかが大きな問題となる．

本研究では，特許分類自動付与及び類似特許文書検索の精度向上方式について検討，評価，考察を行ってきた．しかし，ユーザの要求する精度には未だに届いておらず，上述したように問題は山積みである．本研究で進めてきた検索精度向上の継続に加え，検索を支援する機能の拡充及び検索システムの使い勝手の向上を含めた利用者支援の観点からも研究を並行して進めていく必要があろう．

謝　辞

本論文の主査である名古屋大学大学院情報科学研究科メディア科学専攻の大西昇教授，副査である名古屋大学情報メディア教育センターの長尾確教授，名古屋大学大学院情報科学研究科メディア科学専攻の村瀬洋教授には，本論文をまとめるにあたってご指導を賜り，厚く御礼申し上げます．

筆者の文書処理研究は，1987 年名古屋大学工学部電気工学科 4 年の時に始まりました．筆者が文書処理研究を始める機会を与えて下さり，以来御指導頂きました名古屋大学名誉教授の杉江昇先生に心から御礼申し上げます．

平成 2 年 4 月に株式会社日立製作所システム開発研究所に入社してからも自然言語処理，文書処理の研究を継続的に行ってきており，これまでに多くの方々から御意見・御指導を賜りました．入社以来，これらの研究開発に従事する機会を与えて下さった，株式会社日立製作所システム開発研究所元所長堂免信義氏，同研究所元所長春名公一博士，同研究所元所長片岡雅憲氏，同研究所元所長和歌森文男博士，同研究所前所長小坂満隆博士，同研究所現所長前田章博士に深く感謝致します．また，筆者の上司として長くお世話になった絹川博之博士，矢島敬士博士（ともに現在，東京電機大学教授），辻洋博士（現在，大阪府立大学教授）には，本研究全般に亘って御支援，御指導を賜りました．深く感謝致します．また，筆者の同僚である平井千秋氏，難波康晴博士，森本由起子氏には，本研究全般に亘って活発な技術討論を戴き，研究の方向付けに役立てることが出来ました．感謝致します．

本論文の各章の内容に関して，多くの貴重な御意見を頂きました．

第 2 章については，日本自転車振興会からの補助金を原資として，財団法人工業所有権協力センターからの委託研究という形で特許分類付与技術の研究を行いました．本研究開発に従事する機会を与えて下さった財団法人工業所有権協力センターの石原正博氏をはじめとする方々に御礼申し上げます．また当時，本研究を共同で推進するとともに，事業の観点から多くの御意見を戴いた株式会社日立製作所公共システム事業部細矢良智氏，木山忠博氏，甲谷和也氏，藤田恵美理氏，同社公共システム営業統括本部大越信幸氏に厚く御礼申し上げます．

第 3 章については，類似特許文書検索技術の研究開発に従事する機会を与えて下さり，特許文書処理全般に亘って御意見戴いた株式会社日立製作所公共システム事業部平川真氏，細矢良智氏，濱川雅之氏，内木良彰氏，久連石一毅氏に深く感謝致します．また，本研究を共同で推進し，技術的な議論を戴いた同社中央研究所岩山真博士，森本康嗣氏，秋

良直人氏，同社ソフトウェア事業部多田勝己氏，松林忠孝氏，小川祐一氏，弥生隆明氏，同社知的財産権本部の西山和博氏に深く感謝致します．

第 4 章については，岩山真博士，絹川博之博士，矢島敬士博士，辻洋博士から有益な御助言を頂きました．深謝致します．

本研究は，名古屋大学大学院情報科学研究科メディア科学専攻知能メディア工学講座の大西研究室のセミナーで発表する機会を与えられ，大西昇教授，工藤博章助教授，竹内義則助教授，松本哲也助手をはじめとして，大学院生及び学部生の方々から熱心な御討論・御意見を賜りました．感謝致します．

最後に，筆者が本論文の執筆を開始するにあたり，御指導及び激励を戴きました株式会社日立製作所システム開発研究所主管研究長舩橋誠壽博士，同研究所田中厚氏，真野宏之氏に厚く御礼申し上げます．

参考文献

[1] 相澤：低頻度語の利用によるテキスト分類性能の改善と評価，情報処理学会論文誌 Vol.44，No.7，pp.1720-1730（2003）.

[2] アンソニー・ケニー，吉岡訳:文章の計量 -文学研究のための計量文体学入門-，南雲堂（1995）.

[3] 荒谷，藤田，菅原：ウェブページ間相互評価によるウェブ検索手法の提案と実装評価，情報処理学会論文誌，Vol.46，No.2，pp.337-347（2005）.

[4] Bao Y.，Du X. and Ishii N.：Improving Performance of the k-Nearest Neighbor Classifier by Combining Feature Selection with Feature Weighting，人工知能学会論文誌，Vol.17，No.3，pp.209-216（2002）.

[5] Bear J.，Israel D.，Petit J. and Martin D.：Using Information Extraction to Improve Document Retrieval，The Sixth Text Retrieval Conference (TREC-6)，pp.367-378（1997）.

[6] 別所，岩瀬，戸部，福村：自然言語検索システムにおける分野推論方式，電子情報通信学会論文誌，Vol.J81-DⅡ，No.6，pp.1317-1327（1998）.

[7] Cai L. and Hofmann T.：Text categorization by boosting automatically extracted concepts，Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval，pp.182-189（2003）.

[8] 陳：新聞の各紙面に見られる文体の類型 -主成分分析法による朝日新聞と読売新聞の分析から-，国語学研究，Vol.42（2003）.

[9] Clarke C. L.，Cormack G.V. and Tudhope E.A.：Relevance Ranking for One to Three Term Queries，In Proceedings of RIAO-97，5th International Conference “Recherche d’Information Assistee par Ordinateur”，pp.388-400（1997）.

[10] Deerwester S.，Dumais T. S.，Furnas W. G.，Landauer K. T. and Harshman R.：Indexing by Latent Semantic Analysis，Journal of the American Society of Information Science，Vol.41，No.6，pp.391-407（1990）.

[11] 土井：文末態度表現に注目した Web Page の調査，情報処理学会研究報告自然言語処理，No.130-7，pp.49-56（1999）.

[12] 藤井，鈴木，辻：段落内共起情報を利用した文書自動分類方式，情報処理学会論文誌 Vol.42，No.3，pp.495-506（2001）.

[13] Fujii A.，Iwayama M. and Kando N.：Overview of Patent Retrieval Task at

NTCIR-4, Proceedings of NTCIR4 Workshop, pp.225-232 (2004).

[14] Fujii A., Iwayama M. and Kando N.: Overview of Patent Retrieval Task at NTCIR-5, Proceedings of NTCIR5 Workshop, pp.269-277 (2005).

[15] Fujii A. and Ishikawa T.: Document Structure Analysis in Associative Patent Retrieval, Proceedings of NTCIR4 Workshop, pp.233-237 (2004).

[16] 藤澤, 絹川：日本語情報処理の諸相　日本語情報検索技術の系譜, 情報処理, Vol.44, No.12, pp.1276-1283 (2003).

[17] 深谷, 山村, 工藤, 松本, 竹内, 大西：単語の頻度統計を用いた文章の類似性の定量化 -部分的類似性の考慮-, 電子情報通信学会論文誌 Vol.J87-D2, No.2, pp.661-672 (2004).

[18] Fukumoto J., Kato T. and Masui F. : Question Answering Challenge for Five ranked answers and List answers - Overview of NTCIR4 QAC2 Subtask 1 and 2 -, NTCIR4 Workshop 4 Meeting, pp.283-290 (2004).

[19] 福本, 鈴木:辞書の語義文を用いた文書の自動分類, 情報処理学会論文誌 Vol.37, No.10, pp.1789-1799 (1996).

[20] 福本, 鈴木：語の重み付け学習を用いた文書の自動分類, 情報処理学会論文誌 Vol.40, No.4, pp.1782-1791 (1999).

[21] 福本,鈴木:WordNet の同義語クラスとその上位関係を利用した文書の自動分類, 情報処理学会論文誌 Vol.43, No.6, pp.1852-1865 (2002).

[22] Fujita S. : Revisiting the Document Length Hypotheses NTCIR-4 CLIR and Patent Experiments at Patolis, Proceedings of NTCIR4 Workshop, pp.225-232 (2004).

[23] 原, 中島, 木谷：テキストのフォーマットと単語の範囲内重要度を利用したキーワード抽出, 情報処理学会論文誌 Vol.38, No.2, pp.299-309 (1997).

[24] Hayes J. and Weinstein S.P. : CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories, Proceedings of Second Annual Conference on Innovative Applications of Artificial Intelligence, pp.1-5 (1990).

[25] Holmes, D.I. : Authorship Attribution, Computers and Humanities, Vol.28, pp.87-106 (1994).

[26] 堀井, 今井, 千原：ディジタル図書館のための概念情報を用いた科学技術論文の検索, 電子情報通信学会論文誌, Vol.J82-DI, No.10, pp.1245-1255 (1999).

[27] 池原, 村上, 木本：単語意味属性を使用したベクトル空間法, 言語処理学会論文誌 Vol.10, No.2, pp.111-128 (2003).

[28] 稲子, 笠原, 湯川, 加藤, 北：概念検索に基づく技術内容からのエキスパートの

検索，情報処理学会論文誌，Vol.45，No.2，pp.614-621（2004）.
[29] 井上，宮崎：Yahoo! Search Technogoly（YST）と，検索分野における Yahoo! JAPAN の戦略，情報処理，Vol.46，No.9，pp.988-994（2005）.
[30] 乾，村田，内元，井佐原：表層表現に着目した自由回答アンケートの意図に基づく自動分類，言語処理学会論文誌 Vol.10，No.2，pp.19-42（2003）.
[31] Itoh H.: NTCIR-4 Patent Retrieval Experiments at RICOH, Proceedings of NTCIR4 Workshop, pp.246-249（2004）.
[32] Iwayama M., Fujii A. and Kando N.: Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task, Proceedings of NTCIR5 Workshop, pp.278-286（2005）.
[33] 亀田，藤崎：テーマ・キー概念・キーワード間の階層構造を利用する新聞記事情報の分類・検索システム，情報処理学会論文誌，Vol.28，No.11，pp.1103-1111（1987）.
[34] Kando N.: Overview of the Fourth NTCIR Workshop, Proceedings of NTCIR4 Workshop, pp.i-viii（2004）.
[35] Kando N.: Overview of the Fifth NTCIR Workshop, Proceedings of NTCIR5 Workshop, pp.i-viii（2005）.
[36] 桂田，小山，大原，馬場口，北橋：文書分類システムの分類誤りに着目した分類ルール修正法，情報処理学会論文誌 Vol.43，No.6，pp.1880-1889（2002）.
[37] 河合：意味属性の学習結果にもとづく文書自動分類方式，情報処理学会論文誌 Vol.33，No.9，pp.1114-1122（1992）.
[38] 河合，塚本，山本，椎野：文書構造を利用した箇条書きや表形式文書からの内容抽出，電子情報通信学会論文誌 Vol.J81-D2，No.7，pp.1609-1620（1998）.
[39] 木本：日本語新聞記事からのキーワード自動抽出と重要度評価，電子情報通信学会論文誌 Vol.J74-D1，No.8，pp.556-566（1991）.
[40] 木本，小川，石川，増永，福島，田中，中渡瀬，芥子，豊浦，宮内，上田，松井，木谷，三池，酒井，徳永，鶴岡，安形：日本語情報検索システム評価用テストコレクションの構築，情報処理学会論文誌，Vol.40，No.9，pp.3537-3553（1999）.
[41] 絹川，木村：日本語文構造解析による自動インデクシング方式，情報処理学会論文誌，Vol.21，No.3，pp.200-207（1980）.
[42] 岸田：情報検索技術とテストコレクション，情報処理，Vol.41，No.8，pp.898-901（2000）.
[43] Kishida K.：Pseudo Relevance Feedback Method based on Taylor Expansion of Retrieval Function in NTCIR-3 Patent Retrieval Task, ACL-2003 Workshop on

Patent Corpus Processing, pp.33-40 (2003).

[44] Konishi K., Kitauchi A., Takaki T.: Invalidity Patent Search System of NTT DATA, Proceedings of NTCIR4 Workshop, pp.250-255 (2004).

[45] 工藤, 平井：議事録を利用した設計レビュー管理システムの開発と評価, 情報処理学会論文誌, Vol.44, No.5, pp.1404-1412 (2003).

[46] 工藤, 松本：半構造化テキストの分類のためのブースティングアルゴリズム, 情報処理学会論文誌, Vol.45, No.9, pp.2146-2156 (2004).

[47] 呉, 山田, 岸本：名詞頻度を使った分類用辞書の構築と評価, 電子情報通信学会論文誌 Vol.J84-D1, No.2, pp.213-221 (2001).

[48] Larkey L.S. : A patent search and classification system, Proceedings of the fourth ACM conference on Digital libraries, pp.179-187 (1999).

[49] Lee Y. and Myaeng S.H. : Text genre classification with genre-revealing and subject-revealing features, Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, pp.145-150 (2002).

[50] Luhn H. P.: A statistical approach to mechanized encoding and searching of literary information, IBM Journal of Research and Development, 1-4, pp.390-417 (1957).

[51] Luhn H. P.: The Automatic Creation of Literature Abstracts, IBM JOURNAL of Research and Development, Vol.2, pp.159-165 (1958).

[52] Maron, M. E.: Automatic Indexing: An Experimental Inquiry, Journal of the Association for Computing Machinery, No.8-3, pp.404-417 (1961).

[53] 間瀬, 絹川, 森井, 中尾, 畑村：思考過程の思考展開図表現に基づく機械設計支援システム, 人工知能学会論文誌, Vol.17, No.1, pp.94-103 (2002).

[54] Mase H., Matsubayashi T., Ogawa Y., Iwayama M. and Oshio T. : Proposal of Two-Stage Patent Retrieval Method Considering Claim Structure, ACM Transactions on Asian Language Information Processing (ACM-TALIP), Vol.4, No.2, pp.186-202 (2005).

[55] Mase H., Matsubayashi T., Ogawa Y., Iwayama M. and Oshio T. : Two-Stage Patent Retrieval Method Considering Claim Structure, NTCIR4 Workshop 4 Meeting, pp. 256-261 (2004).

[56] Mase H., Matsubayashi T., Ogawa Y., Yayoi T., Sato Y. and Iwayama M. : NTCIR5 Patent Retrieval Experiments at Hitachi, Proceedings of NTCIR Workshop 5 Meeting, pp.318-323 (2005).

[57] 間瀬久雄，大西昇：特許文書中のタームの出願人別使用傾向の分析と類似特許文書検索精度への影響評価，情報処理学会デジタルドキュメント研究会，54-11，pp.77-84（2006）.

[58] 間瀬，大西：特許出願人に関する傾向の分析とそれを適用した無効化特許検索手法，電気学会論文誌，Vol.127-C，No.1，pp.44-51（2007）.

[59] 間瀬，辻，絹川，石原：特許テーマ分類方式の提案とその評価実験，情報処理学会論文誌，Vol.39，No.7，pp.2207-2216（1998）.

[60] Mase H.，Tsuji H.，Kinukawa H.，Hosoya Y.，Koutani K. and Kiyota K.：Experimental Simulation for Automatic Patent Categorization，Advanced Production Management System（APMS96），pp.377-382，（1996）.

[61] 松井,難波,井形:全文検索エンジン，情報の科学と技術，50(1)，pp.9-13（2000）.

[62] 松本：形態素解析システム「茶筌」，情報処理，Vol.41，No.11，pp.1208-1214（2000）.

[63] 松尾，石塚：語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム，人工知能学会論文誌 Vol.17，No.3，pp.217-223（2002）.

[64] 水野：類似文献自動検索システムについて，特技懇，No.223（2002）.

[65] 森本，間瀬，平井，衣川：問合せ事例を活用したヘルプデスクオペレータ支援機能の開発，情報処理学会論文誌，Vol.44，No.7，pp.1731-1739（2003）.

[66] 森本，間瀬，平井，阿部，大野：システムエンジニア向け情報共有システムの開発，プロジェクトマネジメント学会誌，Vol.7，No.2，pp.40-45（2005）.

[67] 村上，嶋村，中島：設計支援のための物理量と語句に基づく設計事例の類似度定義，人工知能学会論文誌，Vol.17，No.1，pp.85-93（2002）.

[68] 村上：真贋の科学，計量文献学入門，朝倉書店（1994）.

[69] 村田，馬，内元，小作，内山，井佐原：位置情報と分野情報を用いた情報検索，言語処理学会論文誌 Vol.7，No.2，pp.141-160（2000）.

[70] 長尾：日本語文献における重要語の自動抽出，情報処理学会誌，Vol.17，No.2，pp.110-117（1976）.

[71] 長尾，森：大規模日本語テキストのｎグラム統計の作り方と語句の自動抽出，情報処理学会自然言語処理研究報告，NL-96，pp.1-8（1993）.

[72] 中川，湯本，森：出現頻度と連接頻度に基づく専門用語抽出，言語処理学会論文誌 Vol.10，No.1，pp.27-46（2003）.

[73] 仲川，木下，高田，関：対話的に調整可能な文書ランキング　－ WWW検索支援の一手法 －，情報処理学会論文誌，Vol.43，No.9，pp.2850-2863（2002）.

[74] 中島:シソーラスを用いた語の共起関係推定による Rocchio フィードバックの精

度向上，情報処理学会論文誌 Vol.43，No.5，pp.1457-1469（2002）.

[75] 中島，木谷，岡田：検索語間における共起関係の特定によるレレバンスフィードバックの高精度化，情報処理学会論文誌，Vol.40，No.3，pp.1236-1244（1999）.

[76] 那須川，諸橋，長野：テキストマイニング － 膨大な文書データの自動分析による知識発見 －，情報処理，Vol.40，No.4，pp.358-364（1999）.

[77] Niwa Y.，Nishioka S.，Iwayama M. and Takano A.：Interactive Document Retieval Interface: DualNAVI，Workshop on Lexical Resources for Information Retrieval，Stuttgart，Germany（1998）.

[78] 小川，松田：n-gram 索引を用いた効果的な文書検索法，電子情報通信学会論文誌，Vol.82-DI，No.1，pp.121-129（1999）.

[79] 小川：擬似類似度法：n-gram 索引のための高速な日本語文書のランキング検索法，電子情報通信学会論文誌，Vol.J83-DI，No.10，pp.1043-1054（2000）.

[80] 大井，隅田，飯田：意味的類似性と多義解消を用いた文書検索手法，言語処理学会論文誌 Vol.4，No.3，pp.51-70（1997）.

[81] 大竹，増山，山本：名詞の連接情報を用いた関連文書検索手法，情報処理学会論文誌 Vol.40，No.5，pp.2460-2467（1999）.

[82] 大山，影浦，神門，木村，丸山，吉岡，高橋：大規模学術情報データベースに適した情報検索システムの開発，電子情報通信学会論文誌 Vol.J84-D1，No.6，pp.658-670（2001）.

[83] Page L.，Brin S.，Motwani R. and Winograd T.：The PageRank Citation Ranking: Bringing Order to the Web，Proceedings of the 7th WWW Conference，pp.161-172（1998）.

[84] Pirolli P.，Shank P.，Hearst M. and Diehl C.：Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection，Proceedings of CHI96，pp.213-220（1996）.

[85] Riloff，E.：Little Words Can Make a Big Difference for Text Classification，Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval，pp.130-136（1995）.

[86] Robertson，S. E. and Sparck Jones，K.：Relevance weighting of search terms，Journal of the American Societ for Information Science，27-3，pp.129-146（1976）.

[87] Robertson S.E.，Walker S.，Beaulieu M.，Gatford M. and Payne A.：Okapi at TREC4，Proceedings of TREC-4，pp.73-86（1996）.

[88] Salton G. and Buckley C.：Term-weighting approaches in automatic text

retrieval, Information Processing & Management, 24-5, pp.513-523 (1988).
[89] Salton G. and McGill M. J. : Introduction to Modern Information Retrieval, McGraw-Hill Book Company (1983).
[90] Sarasua L. and Corremans G. : Cross Lingual Issues in Patent Retrieval, In Proceedings of ACM SIGIR 2000 Workshop on Patent Retrieval (2000).
[91] 笹島, 浜野：日本人のための検索技術を目指して －gooにおける日本語検索の取り組みー, 情報処理学会誌, Vol.46, No.9, pp.995-1000 (2005).
[92] Shannon C.E. : A mathematical theory of communication, Bell System Tech.J., Vol.27, pp.379-423, pp.623-656 (1948).
[93] 新森, 奥村, 丸川, 岩山：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).
[94] 新谷, 角田, 大石, 長尾：単語の共起頻度と出現位置による新聞の関連記事の検索手法, 情報処理学会論文誌Vol.38, No.4, pp.855-862 (1997).
[95] 白木, 黒橋：自然言語入力と目次との柔軟な照合による図書検索システム, 情報処理学会論文誌Vol.41, No.4, pp.1162-1170 (2000).
[96] Sparck Jones K.: A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, 28-1, pp.11-21 (1972).
[97] Streeter L.A. and Lochbaum K.E. : Who Knows: A System Based on Automatic Representation of Semantic Structure, Proceedings of RIAO88, pp.380-388 (1988).
[98] 杉山, 波多野, 吉川, 植村：ハイパリンクで結ばれた隣接ページの内容に基づくWebページのためのTF-IDF法の改良,電子情報通信学会論文誌Vol.J87-D1, No.2, pp.113-125 (2004).
[99] Sumner R. G. Jr. and Shaw W. M. Jr. : An Investigation of Relevance Feedback Using Adaptive Linear and Probabilistic Models, The Fifth Text Retrieval Conference (TREC-5) (1997).
[100] 鈴岡, 澤島, 東, 馬瀬：サイトナビゲーション・システム「サイトナビ」, 情報処理学会誌, Vol.46, No.9, pp.1001-1007 (2005).
[101] 平, 春野：Support Vector Machineによるテキスト分類における属性選択, 情報処理学会論文誌, Vol.41, No.4, pp.1113-1123 (2000).
[102] 平, 春野：トランスダクティブ・ブースティング法によるテキスト分類, 情報処理学会論文誌, Vol.43, No.6, pp.1843-1851 (2002).
[103] 高木, 藤井, 石川：検索質問の主題分析に基づく類似文書検索と特許検索への応用, 情報処理学会論文誌, Vol.46, NO.4, pp.1074-1081 (2005).

[104] 高村，松本：文書分類のための共クラスタリング，情報処理学会論文誌 Vol.44, No.2, pp.443-450 (2003).

[105] 高野，西岡，今一，岩山，丹羽，久光，藤尾，徳永，奥村，望月，野本：汎用連想計算エンジンの開発と大規模文書分析への応用，http://geta.ex.nii.ac.jp/pdf/itx2002.pdf (2002).

[106] Takeuchi H., Uramoto N. and Takeda K.: Experiments on Patent Retrieval at NTCIR-4 Workshop, Proceedings of NTCIR4 Workshop, pp.271-275 (2004).

[107] Taylor R. S.: Question-negotiation and information seeking in libraries, College & Research Libraries, 29-3, pp.178-194 (1968).

[108] 徳永：情報検索と言語処理，東京大学出版会 (1999).

[109] 柘植，獅々堀，黒岩，北：サポートベクターマシンによる適合性フィードバックを用いた情報検索，情報処理学会論文誌 Vol.44, No.1, pp.59-67 (2003).

[110] Turtle H. and Croft W. B.: Evaluation of an inference netwoek-based retrieval model, ACM Transactions on Information Systems, 9-3, pp.187-222 (1991).

[111] 上嶋，三浦，塩谷：同義語，多義語の考慮による文書分類の精度向上，電子情報通信学会論文誌 Vol.J87-D1, No.2, pp.137-144 (2004).

[112] 山田，森，中川：複合語マッチングと共起情報を併用する情報検索，情報処理学会論文誌，Vol.39, No.8, pp.2431-2439 (1998).

[113] 山名，近藤：サーチエンジン Google，情報処理，Vol.42, No.8, pp.775-780 (2001).

[114] 山名，村田：検索エンジンの概要，情報処理，Vol.46, No.9, pp.981-987 (2005).

[115] 安本：語彙の量的構造，数理科学 Vol.15, No.6, pp.44-49 (1977).

[116] 余田，湯村，西田：言語情報に基づく検索，Info-Tech94 講演論文集，pp.138-146 (1994).

[117] 吉田：自動大分けシステムから中分けシステム　ー類似文書探索を利用した自動分類付与ツールー，財団法人日本特許情報機構創立 20 周年記念誌　特許情報活用の時代の検索と機械翻訳技術，pp.86-89 (2005).

[118] 湯浅，上田，外川：大量文書データ中の単語間共起を利用した文書分類，情報処理学会論文誌 Vol.36, No.8, pp.1819-1827 (1995).

発表文献リスト

論文

（主筆）

(1) 間瀬久雄，木山忠博，絹川博之：ＤＢ検索用自然語インタフェースにおける解釈結果確認文生成方式の開発とその評価，情報処理学会論文誌，Vol.35，No.8，pp.1579-1590 (1994).

(2) 間瀬久雄，辻洋，絹川博之，石原正博：特許テーマ分類方式の提案とその評価実験，情報処理学会論文誌，Vol.39，No.7，pp.2207-2216 (1998).

(3) 間瀬久雄，辻洋：Experiments on Automatic Web Page Categorization for Information Retrieval System，情報処理学会論文誌，Vol.42，No.2，pp.334-348 (2001).

(4) 間瀬久雄，絹川博之，森井洋，中尾政之，畑村洋太郎：思考過程の思考展開図表現に基づく機械設計支援システム，人工知能学会論文誌，Vol.17，No.1，pp.94-103 (2002).

(5) Mase H.，Matsubayashi T.，Ogawa Y.，Iwayama M. and Oshio T.：Proposal of Two-Stage Patent Retrieval Method Considering Claim Structure，ACM Transactions on Asian Language Information Processing (ACM-TALIP)，Vol.4，No.2，pp.186-202 (2005).

(6) 間瀬久雄，大西昇：特許出願人に関する傾向の分析とそれを適用した無効化特許検索手法，電気学会論文誌，Vol.127-C，No.1，pp.44-51 (2007).

（副筆）

(1) 絹川博之，難波康晴，間瀬久雄，森本由起子，辻洋：自然言語インタフェースＮＬＩ－状況推移ベースモデリングによるＮＬＩ構築システム－，NII Journal，No.4，pp.1-13，国立情報学研究所 (2002).

(2) 森本由起子，間瀬久雄，平井千秋，衣川一久：問合せ事例を活用したヘルプデスクオペレータ支援機能の開発，情報処理学会論文誌，Vol.44，No.7，pp.1731-1739 (2003).

(3) 森本由起子，間瀬久雄，平井千秋，阿部琢哉，大野治：システムエンジニア向け情報共有システムの開発，プロジェクトマネジメント学会誌，Vol.7，No.2，pp.40-45 (2005).

国際会議発表論文

（主筆）

(1) Mase H., Tsuji H. and Kinukawa H. : Computer-Aided News Article Summarization, Ninth International Conference on Industrial and Engineering Applications of AI and Expert Systems (IEA/AIE96), pp.627-632 (1996).

(2) Mase H., Tsuji H., Kinukawa H., Hosoya Y., Koutani K. and Kiyota K. : Experimental Simulation for Automatic Patent Categorization, Advanced Production Management System (APMS96), pp.377-382, (1996).

(3) Mase H., Ishii K., Kinukawa H., Morii H., Nakao M. and Hatamura Y. : Creative Design Support System Based on Designer-Thinking-Process Model, Symposium on Analysis, Design, and Evaluation of Human-Machine Systems (2001).

(4) Mase H., Matsubayashi T., Ogawa Y., Iwayama M. and Oshio T. : Two-Stage Patent Retrieval Method Considering Claim Structure, NTCIR Workshop 4 Meeting, pp. 256-261 (2004).

(5) Mase H., Matsubayashi T., Ogawa Y., Yayoi T., Sato Y. and Iwayama M. : NTCIR5 Patent Retrieval Experiments at Hitachi, NTCIR Workshop 5 Meeting, pp.318-323 (2005).

（副筆）

(1) Tsuji H., Namba Y., Mase H., Morimoto Y. and Kinukawa H. : Building Natural Language Interface - Its Methodology and Tools -, 6th IFAC on Analysis, Design and Evaluation of Man-Machine System, pp.389-394 (1995).

(2) Tsuji H., Morimoto Y., Namba Y., Mase H., Kinukawa H. and Endoh H. : User Guidance Function in Natural Language Interface for Document Query and Handling, Flexible Query-Answering Systems Proceedings of the 1996 Workshop (FQAS96), pp.105-118 (1996).

(3) Tokuda T., Mase H., Tsuji H. and Niwa Y. : Experimental Evaluation on Associative Keyword Reminder by Thesaurus, 7th IFAC Symposium on Analysis, Design and Evaluation of Man-Machine Systems, pp.161-166 (1998).

(4) Morimoto Y., Mase H., Hirai C., Tsuji H. and Kinugawa K. : Operator Navigation System for Help Desk, World Multiconference on Systemics, Cybernetics and Informatics, pp.173-178 (2000).

(5) Morimoto Y., Mase H. and Hirai C. : Development and Evaluation of a Reuse Process Support System for Document-type Knowledge, ProMac2006, CD-ROM (2006).

国内雑誌寄稿

（主筆）

(1) 間瀬久雄, 岩山真, 松林忠孝, 小川祐一, 大塩只明：文章構造を利用した二段階特許検索方式の提案と評価, 財団法人日本特許情報機構創立 20 周年記念誌　特許情報活用の時代の検索と機械翻訳技術, pp.96-101 (2005).

外国雑誌寄稿

（主筆）

(1) Mase H., Morimoto Y., Tsuji H. and Kinukawa H. : Classification Knowledge Discovery From Newspaper Ariticles, Studies in Informatics and Control, Vol.9, No.3, pp.167-178 (2000).

テクニカルレポート

（主筆）

(1) Mase H. : Experiments on Automatic Web Page Categorization for IR system, Stanford University Computer Science Department Database Group Technical Report, 1998-18 (1998).

（副筆）

(1) Tsuji H., Morimoto Y., Namba Y., Mase H., Kinukawa H. and Endoh H.：User Guidance Function in Natural Language Interface for Document Query and Handling, In Technical Report No.65 of Roskilde University, pp.105-118 (1996).

国内研究会・シンポジウム発表論文

（主筆）

(1) 間瀬久雄，大西昇，杉江昇：説明文の抄録作成について，電子情報通信学会言語理解とコミュニケーション研究会，NLC89-40，pp. 5-12（1990）.

(2) 間瀬久雄, 辻洋, 絹川博之:新聞記事要約作成支援システムのユーザインタフェース，計測自動制御学会第 10 回ヒューマンインタフェース・シンポジウム，pp. 147-150（1994）.

(3) 間瀬久雄，大西昇：特許文書中のタームの出願人別使用傾向の分析と類似特許文書検索精度への影響評価，情報処理学会デジタルドキュメント研究会，54-11，pp. 77-84（2006）.

（副筆）

(1) 辻洋，難波康晴，間瀬久雄，森本由起子，絹川博之：自然語インタフェースの構築：方法論と応用，情報処理学会ヒューマンインタフェース研究会，57-4，pp. 25-32（1994）.

(2) 杉本雅則，小山照夫，堀浩一，大須賀節雄，絹川博之，間瀬久雄：文書間の関連性を可視化することによる文献検索システム，情報処理学会自然言語処理研究会，NL112，pp. 15-22（1996）.

(3) 辻洋，間瀬久雄，津原進，衣川一久：ヘルプデスクにおける類似文書検索システムの構成と機能について，情報処理学会デジタルドキュメント研究会，10-4，pp. 23-30（1997）.

(4) 森本由起子，間瀬久雄，辻洋：新聞データからの分類知識獲得に関する実験シミュレーション，情報処理学会デジタルドキュメント研究会，6-1，pp. 1-8（1997）.

(5) 徳田圭世，間瀬久雄，辻洋：デジタルドキュメントにおける共起データを用いた検索タームの連想支援について，情報処理学会デジタルドキュメント研究会，10-3，pp. 15-22（1997）.

(6) 森本由起子，間瀬久雄，辻洋，衣川一久：ヘルプデスクシステムにおける類似事例検索機能の開発及び評価，情報処理学会デジタルドキュメント研究会，20-1，pp. 1-8（1999）.

(7) 森本由起子，間瀬久雄，水野浩孝，辻洋，遠藤武之：単語出現頻度に基づくテキスト分類ツールとその応用，日本計算機学会第 13 回大会，pp. 76-77（1999）.

(8) 森本由起子，間瀬久雄，平井千秋，阿部琢哉，大野治：システムエンジニア向け情報共有システムの開発，プロジェクトマネジメント学会春季研究発表大会予稿集，pp.173-178（2004）.

(9) 森本由起子，間瀬久雄，平井千秋，辻洋：文書化知識の再利用過程の支援システムに関する考察，電気学会情報システム研究会，電気学会第 25 回情報システム研究会，IS-06-11，pp.61-66（2006）.

全国大会発表論文

（主筆）

(1) 間瀬久雄，木山忠博，絹川博之：自然語インタフェースにおけるインタラクティブ型多義解消方式の開発，情報処理学会第 44 回全国大会，Vol.3，pp.235-236（1992）.

(2) 間瀬久雄，木山忠博，辻洋，絹川博之：自然語インタフェースにおける解釈結果確認文生成方式の開発，情報処理学会第 45 回全国大会，Vol.3，pp.135-136（1992）.

(3) 間瀬久雄，辻洋，川村隆雄，絹川博之：形態素解析ツールによるかな漢字プログラミングの実現，情報処理学会第 46 回全国大会，Vol.3，pp.127-128（1993）.

(4) 間瀬久雄，小山幸子，木山忠博，辻洋，絹川博之：文字認識と形態素解析を用いた類似文書検索の試み，情報処理学会第 47 回全国大会，Vol.3，pp.161-162（1993）.

(5) 間瀬久雄，辻洋，絹川博之：パラメータ設定による文章要約支援システム，情報処理学会第 48 回全国大会，Vol.3，pp.103-104（1994）.

(6) 間瀬久雄，森本由起子，辻洋，絹川博之：テキスト分類支援ツールＦＬＵＴＥの開発（1）－ 機能と構成 －，情報処理学会第 52 回全国大会，Vol.3，pp.303-304（1996）.

（副筆）

(1) 辻洋，間瀬久雄，木山忠博，絹川博之：テキスト自動分類エキスパートシステムの一構成法，第 49 回情報処理学会全国大会論文集，3J-8（1994）.

(2) 森本，間瀬久雄，辻洋，絹川博之：テキスト分類支援ツールＦＬＵＴＥの開発(2)－障害事例分類への適用 －，情報処理学会第 52 回全国大会，Vol.3，pp.305-306（1996）.

(3) 徳田圭世，西川記史，辻洋，間瀬久雄，丹羽芳樹：情報検索発想支援のためのシソーラス管理システムの提案,情報処理学会第 53 回全国大会，Vol.3，pp.163-164（1996）.

(4) 森本由起子，間瀬久雄，辻洋，絹川博之：新聞記事自動分類システム構築の検討と評価，情報処理学会第 53 回全国大会，Vol.3，pp.205-206（1996）.

(5) 徳田圭世，西川記史，細川貴史，間瀬久雄，辻：シソーラス管理システムにおけるカスタマイズ機能について,情報処理学会第 54 回全国大会, Vol.3, pp.171-172 (1997).

(6) 徳田圭世，間瀬久雄，森本由起子，辻洋，丹羽芳樹：WWWホームページからの共起語自動抽出実験，情報処理学会第 55 回全国大会，Vol.3，pp.72-73 (1997).