学位報告

報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　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

氏　　名　間瀬　久雄

論文内容の要旨

本論文は，特許文書を対象とした文書分類・検索の精度向上技術に関する研究成果をまとめたものである．

まず，本研究の背景について述べる．

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，特許の戦略的な取得・管理・活用に力を入れるようになっている．その一方で，特許庁における特許審査請求件数はこの数年で急増しており，審査請求から審査開始までの審査待ち期間が26ヶ月にもなっている．その結果，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，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．

次に，本研究の目的について述べる．

特許審査業務では，出願特許の発明内容に応じて特許分類を割り当てる分類付与作業と，出願特許の発明内容を無効化する過去の文献を検索する類似特許検索作業に多大な時間がかかっている．そこで，本研究の目的は，上記背景を鑑み，特許分類自動

付与精度及び類似特許文書検索精度を向上させることにより，審査業務の効率改善と品質の向上を実現することにある．

次に，本研究の基本課題について述べる．

本研究における分類自動付与では，出願特許文書を入力として，既存の特許分類体系のうち，発明内容に合致した適切な分類を自動付与する．また，類似特許文書検索では，発明内容を端的に記載している請求項１の文章を入力として，その発明内容を無効化する過去の特許文書を検索する．本研究では，従来研究と同様，特許文書中のタームの出現傾向を用いるというアプローチを前提とする．その際，(1) 特許文書に記載される発明内容をどのように定式化するか（発明内容を特徴付けるタームをどのようにして特定するか），(2) 特定されたタームを用いて，文書と分類，文書と文書の間の類似性をどのように算出するか，という精度向上を実現するための基本課題を解決する必要がある．

次に，本研究の特徴について述べる．

本研究では，上記基本課題を解決するために，特許文書の記載方法に関する特徴を踏まえ，特許文書内の言語構造，すなわち，特許文書の構成及び特許文書に記載される文章の構文的・語彙的特徴を利用して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って発明内容を定式化し，文書間（または文書と分類の間）の類似度を算出する方式を提案し，その有効性を評価している．

本論文の構成と内容は以下の通りである．

第 1 章では，はじめに特許を巡る社会動向および特許庁での特許審査に係る課題について述べている．特許審査では，出願特許に対して特許分類を付与する作業と，類似する特許文書を検索する作業の効率化が課題となっている．次にこれらの効率化を計算機を用いて実現するための基本課題と，これらの課題を解決するために本研究が

採ったアプローチについて述べている.

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ている．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ている．ここでは，発明内容を端的に記載した請求項１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している．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，本研究と比較している．まず，文書検索の歴史的経緯について簡単に触れている．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している．次に，分類自動付与に係る研究動向について述べている．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している．

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ている．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している．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．