拒絶理由を用いた特許文献と先行技術の類似性判定に関する研究

筑波大学審査学位論文（博士）

2015

柳堀　恭子

筑波大学大学院
ビジネス科学研究科　企業科学専攻
システムズ・マネジメントコース

拒絶理由を用いた特許文献と先行技術の類似性判定に関する研究

目次

論文概要

博士（システムズ・マネジメント）

拒絶理由を用いた特許文献と先行技術の類似性判定に関する研究

筑波大学大学院
ビジネス科学研究科　企業科学専攻
システムズ・マネジメントコース
柳堀　恭子

特許庁が 2001 年に打ち出した出願から審査着手までの期間を短縮するという施策の効果などより，我が国の特許査定率は上昇傾向にある．しかし，依然として拒絶査定は全出願の半数近くにのぼっている．

そして，出願から審査着手までの期間が短くなることで，審査官が，文献を十分に検索することができる可能性が低下することが予見できる．一方，出願人についても特許出願を厳選するという傾向が見られる．これは，主たる出願者である各企業の研究開発費の減少，あるいは知財費用の減少など業績の悪化を原因とするもの，先行する技術調査の強化による質の高い特許を目指し出願

を厳選したとするもの，出願のノルマ制廃止・企業戦略上必要なもののみ出願するといった企業内の知財活動の変化がもたらした結果といえる．

このような情勢を鑑みると，審査官のみならず，出願人の立場からみた場合も，特許権を得るために，限られた時間の中での適切に先行技術調査を行うことが重要であることがわかる．

出願人が先行技術調査を行う目的は，出願内容と酷似する文書，類似する文書，同じであると推定されるような文書を確認するためである．出願前に類似特許を発見した場合，出願する特許の新規性は否定されることになる．そのため，多くの場合，審査請求時に拒絶査定を受ける可能性が高く，仮に拒絶されず特許になったとしても，その後，特許紛争になる懸念もある．そのために先行技術調査を行うのである．

文書同士を比較し，類似文書と定義するための方法は，元来より存在している．いずれも文中にスペースのない日本語文に対して，単語ごとに区切り，その品詞を同定する形態素解析という手段を利用する方法をベースとしている．対象文書を形態素解析し，得られた単語とその頻度，単語の重要度から対象文書と同じような単語の頻度や重要度をもつ別の文書を類似文書とする方法である．

しかしながら，特許公報や特許文献（以下，本論では特許文書とする）は，慣習となっている特徴的な記載方法により非常に難読かつ理解し難い文書になってしまっている．そのため，通常の文書解析の方法では文書の類似性を正確に捉えることができない．その理由として，文中に句点が存在せず長い修飾により冗長となっている点，説明の不明瞭さをなくすために，指示代名詞の代わりに使われる「前記」「該」などの接頭辞が頻出する点，出願者が発明の新規性を主張するために，意図的に名詞を組み合わせて造語し，既存の辞書にないような複合名詞を作りだし，その複合名詞が，権利範囲拡大を狙うために限定的ではなく広い概念でつくられているために分かりにくい点などがある．これらが組み合わさり，文書を難解なものにさせている．それにより，人の目で判断する場合においても，どの文書同士が類似であるかという判断をも難しくしてしまっている．

また，類似文書検索をする際に，文中の複合名詞が形態素解析で分解されて

しまうと，複合名詞が本来持つ意味が消滅してしまうことになる．そのうえ，特許特有の語でもある「装置」「特徴」「システム」「手段」などいずれの語が文書中に確認されると，少なからず，比較文書同士が類似する結果になってしまう．しかし，もともと出現数が少ない複合名詞をそのまま利用すると，反対に類似文書が見つかりにくくなってしまうという問題が生じる．

この問題に対して，複合名詞そのものの存在を増やすことができればよいと考えられる．そのため，同じではないが，類似すると思われる複合名詞の候補を増やすため，類似複合名詞のデータを収集する方法が有効であると考えた．複合名詞の収集を考えたときに，通常の辞書では，新語が加わるまでのタイムラグがあるため，新たに辞書構築をする必要があることはわかった．そして，いままでの先行研究で，出願書類の解析は多く目にしてきたが，審査官が示す審査書類の解析については，ほぼ研究例がないということに気づき，この審査書類を解析することとした．これにより審査書類の１つである「拒絶理由通知書」の中に複合名詞が多く記載されている箇所を発見することができた．

本研究では，特許審査の過程で示される「拒絶理由通知書」に着目し，その中で，審査官が，対象文書とその引用の関係で類似箇所として対比している部分の複合名詞対を抽出することで類似複合名詞辞書データを作成し，複合名詞の数を増やし検索に応用することを提案した．

本論文では，特許文書と自然言語処理との関係から，一般的な文書と特許文書の文書類似検索の違いについて述べ，特許文書での解析に必要な課題について述べる．

次に，特許審査官が示す拒絶理由通知書の利用について，その基礎となる特許制度および特許分類について説明する．

つづいて，拒絶理由通知書からの複合名詞抽出とその効果について，形態素解析の手法を用いた文書類似の手法と複合名詞解析の結果を比較することから複合名詞利用の効果を示す．特許文書を対象として形態素解析を行うと，ほぼすべての文書が類似を示してしまい，ノイズが非常に多いという結果より，特許文書の類似分析には，複合名詞を用いることを検討した．しかし，複合名詞解析にも，出現数の少なさから類似文書検索が難しいという問題があり，この問題を考える必要がある．

類似する複合名詞の数が増やすために，類似複合名詞の辞書データを作成する必要がある．けれども，この辞書データの作成には多大な費用と時間を要する点, 当該分野に不明な人が作成すると, 辞書データに誤りが含まれる点など，様々な課題がある．そこで辞書データの作成において，特許の判定を行う専門の審査官が作成した拒絶理由通知書の記載中の対比箇所から抽出することができることを明らかにした.

しかしながら，抽出した複合名詞をすべて類似複合名詞データとして使うことは難しい. なぜなら, その文書だけに固有の複合名詞も存在するからである. そこで，複合名詞自体の類似判定方法について，構文解析手法の利用から判定する方法と複合名詞内形態素の類似から判定する方法について提案した．そして，もともと類似であるとされている文書同士の類似度について，作成した辞書データが機能するかどうかを確認した．まず，形態素解析をベースとし，コサイン類似度で表される文書間の類似と，作成した文書間類似度式を用いた複合名詞と動詞のフレーズで文書間の類似を測定した．その際に作成した辞書データを適用させた．その結果を比較し，本手法を用いることにより，コサイン類似度が低かった文書対でも高い類似性レベルを得られ，類似性判断をより正確に行えるようになったといえた.

さらに, 辞書データを応用し, 複合名詞と動詞のフレーズの類似度を利用し，このフレーズをより上位概念の複合名詞へ集約していくことで文書の類似度数の変化を確認した．この結果，集約をすることで類似複合名詞の群が大きくなり，　類似する文書類似数も増加することがわかった.

拒絶理由通知書から複合名詞を抽出する利点として，審査は特許庁で常に行われており，一定の数は拒絶査定を受け，その際に拒絶理由通知書が必ず示される. そのため, 拒絶理由通知書のデータは日々蓄積され続けており, 我々は，それを利用することができる．そして，それらデータは特許審査官という信頼度の高い質の高いデータを収集することができることと考えられる.

また，時によっては，利用される複合名詞が陳腐化する怖れがあるが，刻々と蓄積されてきた拒絶理由通知書のデータにおいて，その複合名詞が使われていた時代背景をくみ取り，その時代にだけ使われていた複合名詞だとしても，その複合名詞に対して類似とされていた別の複合名詞を判定することもできる

と考えられる．

これらより，文書の類似判定に利用するために，拒絶理由通知書から類似複合名詞の情報を抽出し，辞書データ化することが文書間類似を判定する手段として有効であるということといえる．

第１章

緒論

近年，我が国の特許査定率は上昇傾向にある[Tokkyocho 14]．特許庁の調査によるとその理由は，2001 年の審査請求期間を出願日から 7 年以内から 3 年以内へと変更したことにより特許出願から審査着手までの期間が短縮していること，出願人による特許出願の厳選が進んでいることがあげられている．

出願から審査着手までの期間の短縮については，出願人の早期権利化へのニーズの高まりによるものだが，出願人の早期権利化ニーズへの対応と，適切なサーチとはトレードオフの関係にあることが示されている．審査官が適切なサーチをするためには十分な時間が必要になるが，審査着手までの期間が短くなることで，文献を十分に検索することができる可能性が低下することが予見できる．一方，出願人による特許出願の厳選であるが，我が国の特許出願は，2008 年から 2009 年にかけて，リーマンショックをまたいで約 10%減少している．これに伴い，審査請求件数も同様に 10%減となっている．これは，主たる出願者である企業の業績悪化による研究開発費の減少や知財費用の減少などが考えられる．この対応として，先行技術調査の強化による質の高い特許を目指した出願の厳選，出願のノルマ制廃止，企業戦略上必要なもののみ出願，といった企業内の知財活動の変化がもたらした結果といえる．

このような情勢を鑑みると，審査官のみならず出願人の立場からみたとしても特許権を得るために，限られた時間の中での適切なサーチ（以下，先行技術調査という）を行うことの重要性が浮き彫りとなる．

出願人が先行技術調査を行うため目的は，出願内容と酷似するあるいは類似する特許公報や特許文献（以下，本論では特許文書とする），同じであると推定

されるような過去の特許文書などの類似特許文書（ここでは類似文書と統一する）を確認するためである．出願前に類似文書を発見した場合には，出願発明の権利化のため，記載内容の検討などの対応をしなければならない．このような出願前の類似文書発見のために先行技術調査を行うのである．

類似文書と定義するための方法は，いくつか存在している．いずれも文の中にスペースのない日本語文に対して，単語に区切り，その品詞を同定する形態素解析という手段を利用する方法である．対象文書に対して形態素解析を行い，得られた単語とその頻度，単語の重要度などから文書の特徴情報を生成し，対象文書と同じような文書の特徴情報を持つ文書を類似文書とする方法である．これらは新聞や定型的に書かれている文書については高い精度での検索結果をもたらす[Inui 06]．

しかしながら，本研究が，対象とする特許出願の文書では，慣習となっている特徴的な記載方法により，非常に難読で理解しがたい文書になっている．具体的にいうと，文中に句点が存在せず長い修飾により冗長となっている点，説明の不明瞭さをなくすために使われない指示代名詞の代わりに使われる「前記」「該」などの接頭辞が頻出する点，出願者が発明の新規性を主張するために，意図的に多用する名詞を組み合わせて造語し，既存の辞書にない複合名詞を作りだす点などである．これらの記載方法により，類似文書を検索する前段階の対象文書の理解の時点で問題が生じている．これらの課題を解決する手法として，冗長な文を構造化し見やすくする手法[Shinmori 04][Konishi 06]などがある．

また，文中の複合名詞が形態素解析で分解されてしまうと，複合名詞が本来持つ意味が消滅してしまうことになる．さらには，特許文書中に高い頻度で出現する「装置」「特徴」「システム」「手段」などの語により，比較文書同士が類似すると判定されることがある．それゆえ，形態素に分解されてしまった名詞ではなく複合名詞のまま類似性評価に使用することが望ましい．

けれども，複合名詞の類似性を正確に把握する手法は現状では存在していない．それゆえ，類似する複合名詞を辞書化する作業は，最終的には人手で編纂しているのが現状である．人手による編纂では，多大な費用が必要になると共に，編纂する人のスキル不足などにより誤った内容が辞書に登録されるなどの

課題がある．このような課題に対して，本研究では，これまで着目されていなかった特許審査の過程で示される「拒絶理由通知書」に着目した.「拒絶理由通知書」とは，特許庁の審査官が，対象文書とその引用の関係で類似箇所として対比している複合名詞箇所を抽出する．それゆえ,「拒絶理由通知書」に記載された複合名詞は，専門家の判断を超えた審査結果の判定であり，特許においては，この内容より正確な複合名詞の類似性は存在しない極めて信頼性の高い情報である．さらに，特許審査は永続的であると共に新しい技術に対して実施される．そのため,「拒絶理由通知書」に記載された複合名詞は，常に増加すると共に技術的な進化を反映した情報となる．そこで，本研究では，類似複合名詞辞書データの作成に,「拒絶理由通知書」から情報を抽出し，応用する手法を提案した.

本論文では，まず第 2 章にて，特許文書と自然言語処理との関係について，通常の文書と特許文書の違いについて述べ，特許文書での解析に必要な課題について述べる.

次に，第 3 章にて，拒絶理由通知書の利用について，その基礎となる特許制度および特許分類について説明する．そして，拒絶理由通知書を利用して編纂した辞書を利用することによって，検索において，どのような効果を生じるのか，その有益性について述べる．つづいて，第 4 章にて，拒絶理由通知書からの複合名詞抽出とその効果について述べ，なぜ通常の文書間類似度を計算する方法で特許文書の類似度が効果的に測定できないのかという理由を明らかにし,複合名詞解析を利用する場合，その問題点と複合名詞抽出の方法について説明する．第 5 章では，複合名詞自体の類似判定方法について，構文解析手法の利用から判定する方法と複合名詞内形態素の類似から判定する方法について述べ,複合名詞の類似評価を行う．そして，第 6 章では，複合名詞を利用した文書類似判定について述べ, 作成した辞書データを用いた際の文書間類似度を測定し,さらに辞書データを応用し，複合名詞と動詞のフレーズの類似度を利用し，フレーズ集約することで文書の類似度数の変化を確認した．最後に第 7 章にて結論を述べる.

第２章

特許文書と自然言語処理との関係

本章では，自然言語処理による文章の解析，特に特許文書における解析方法について，これまでの先行研究を概観し，結言において，特許文書検索の課題について示す.

2.1　自然言語処理による文書解析

本節では，自然言語処理，とりわけ一般的なテキストマイニング[Watanabe 03][Nasukawa 09]，分類の技術について述べる.

そのうえで，次節で扱う特許文書解析との違いを明確にする.

2.1.1　形態素解析

テキストから必要な情報をとりだすテキストマイニングを行ううえで必要となり，一般的な文書解析の基礎となっているのが，形態素解析[Yasuda 06]という技術である.

形態素解析を行い，文書を単語に分解し，単語の出現頻度や，対象文書内で

の重要度を単語に重み付けして計算する．そして，類似文書検索では，このように計算された単語情報をもとに，同じような単語情報をもつ文書を探し出す仕組みとなっている．

形態素解析で重要な技術は，解析のアルゴリズムと形態素解析用の辞書の存在である[Hashimoto 12]．辞書をもとにして，どの箇所で単語を区切るかが決定されるのである．

そして，解析は主に 3 つの処理からなる．

事前に辞書や規則データを用意し，文を入力する[Nagao 78]．そうすると，まず，分かち書きという文を単語単位の文字列に区切る処理が行われ，次に活用語の語尾処理を行い原型と活用形などの情報を得る処理，そして個々の単語の品詞を同定する処理が行われていく．

「今日の天気は晴れです」を入力文とした場合の形態素解析のながれを以下の図に示す．

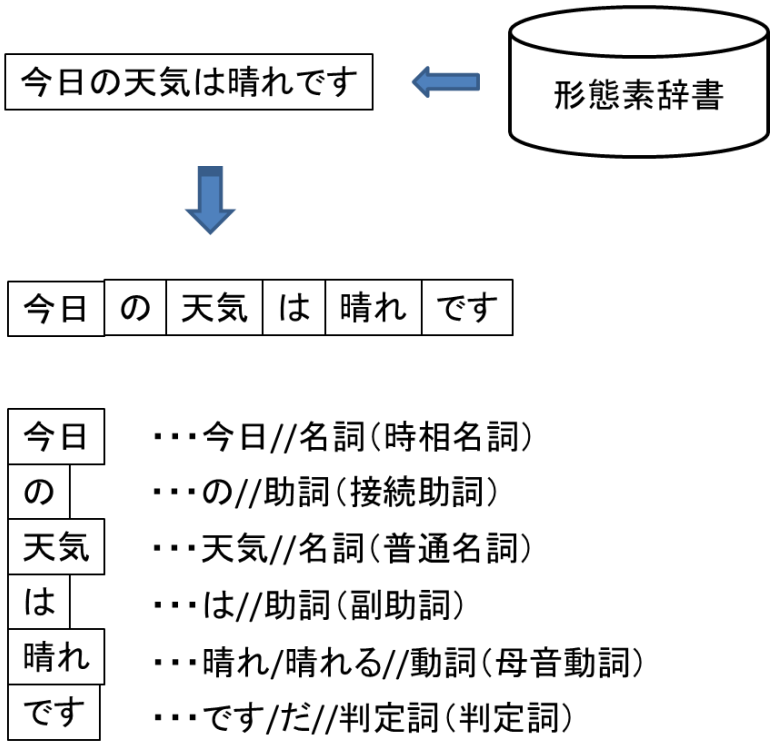


図 2.1　形態素解析の仕組み

図 2.1 のように，入力文を，辞書をもとに最小の形態素ごとに分割し，分割した形態素の語尾処理を行い，各形態素の品詞情報を付す.
このときに，各形態素解析技術 MeCab[Kudo 01]や Chasen[Matsumoto 00][Matsumoto 03]で使用される辞書の違いにより，分割される箇所が違う場合もある[Asahara 03a][Asahara 03b].

形態素解析の技術は，新聞記事のような定型的な書かれ方をしている文書には有効である．しかし，問題点もいくつかある.

まず，形態素解析は既存の辞書を基にしているため，辞書にない語は未知語として判別ができないことである．近年の SNS（Social Networking Service）CGM（Consumer Generated Media）など誰でも発信者となりうる文書については，定型的な書式といったものはなく，使われる語も多様であり，既存の辞書になかったような下記例のような未知語や新語が多数存在する.

例・ディする（disrespect の略，批判するという意味）
・草食系男子（恋愛に縁がないわけではないのに積極的ではない男子）
・婚活する（結婚活動の略）
・ドN（いたって普通でノーマルな人のこと）

このように，時代の流れとともに今まで見聞きしたことのない語が次々と出現してくるため，辞書化が追いついていかない.

未知語や新語の問題に対しては，新しい語が出現するたびに 1 つずつ辞書に追加していくという方法もあるが，手間もかかり効率的ではないうえに，完全に漏れがなくなるわけではない．この問題を解決するのに JUMAN7.0[Juman 12]では，新語などの処理を人手ではなく，コーパスから自動的獲得して構築した辞書を利用している[Murawaki 10].

他にも，単語をどこで区切るかは形態素解析手段が使用する辞書に依存するため, 人が意図する単語の区切りと違った解答となる場合がある. このような,

分かち書きの曖昧性とよばれる問題も存在する[kudo 05].

形態素解析を利用したテキストマイニングの応用研究として，金融，医療などの分野でも活用されている[Izumi 10][Okabe 06][Kimura 05].

2.1.2　文書分類

文書分類とは，その内容に基づいて，文書群を 1 つ以上の群に分類することである．文書分類には，正しい分類に関する情報を与える学習データあり（教師ありデータ）文書分類と，外部の情報を参照せずに分類する学習データなし（教師なしデータ）文書分類とにわけられる．分類を行うことで，大規模な情報の中から有益な情報のみを取り出すことも可能であり，文書分類を応用した研究もされている[Inoue 01][Sasaki 04].

本項では，一般的に使われる文書分類の方法について述べる．

(1) ナイーブベイズ法

ベイズの定理では，ある文書 D がカテゴリ C に分類される確率を P(C | D)で表される．逆に，カテゴリ C の中に文書 D が含まれている確率は P(D | C)で表される．

つまり，文書 D が得られ，カテゴリ C に含まれる確率 P(CD)は式 2.1 に示される．

$$P(CD) = P(C|D)P(D) = P(D|C)P(C) \quad (式\ 2.1)$$

つまり，下記式 2.2 に変形できる．

$$P(C|D) = \frac{P(D|C)P(C)}{P(D)} \quad (式\ 2.2)$$

ここで，P(D | C)は，次式 2.3 で表される．

$$P(D|C) = \frac{\text{文書}D\text{の数}}{\text{カテゴリ}C\text{の文書数}} \qquad \text{(式 2.3)}$$

しかし，同じ文書が見つけるという確率を考えるとすると，現実的にはまったく同じ文書 D が出現する確率は少ない．文書 D の数が 0 になると，P(D | C) も 0 になってしまう．そこで，ナイーブベイズでは，ベイズ定理をもとに，文書を単語群として捉えることで，単語があるカテゴリ C に分類される確率を表すことができる[McCallum 98a].
つまり，カテゴリを特徴づける語多く含むほど，そのカテゴリに分類されやすくなる.

文書 D = 語 W $(\sum_{i=1}^{n} Wi)$

と考えると，

$$P(W|C) = \frac{\text{単語}Wi\text{の数}}{\text{カテゴリ}C\text{の語彙の数}} \qquad \text{(式 2.4)}$$

となる.
カテゴリ C に含まれている Wi の確率は，式 2.5 で表される.

$$\prod P(wi|C)^{n(wi|D)} \qquad \text{(式 2.5)}$$

ここで，$n(wi|D)$ は，分類しようとする文書 D 内の Wi の頻度とする.
よって，文書 D がカテゴリ C に分類される確率を単語 Wi で表すと，式 2.6 となる.

$$P(C|D) \cong P(C) \prod P(wi|C)^{n(wi|D)} \qquad \text{(式 2.6)}$$

文書内に各語がそれぞれ独立に現れるという仮定に基づいているが，現実の

自然言語にはあり得ない仮定なので「ナイーブ」とよばれている．
ナイーブベイズ法は高精度というわけではないが，文書自動分類のような文書のトピックを決めるような場合に使われたりする[Kondo 07][Hanai 05][Hirano 05].

ただし，あるカテゴリで一度も出現したことのない語が現れた場合，確率P(W|C)が 0 になってしまう問題がある．このようなデータに対して，ある程度の確率値を割り当てておき，確率が 0 になることを防ぐ操作のことをスムージングという[Kita 99].

(2) SVM

サポートベクトルマシンという教師あり分類学習の 1 つで，主に 2 値分類に使われる．文書分類では，語を属性としたときに，この属性をベクトルで表現し，このベクトル空間上にカテゴリ面の境界を自動的に学習する．TF･IDF 法による重み付けの値を属性の値にすることが多い[Kurita 03] [Hearst 98].

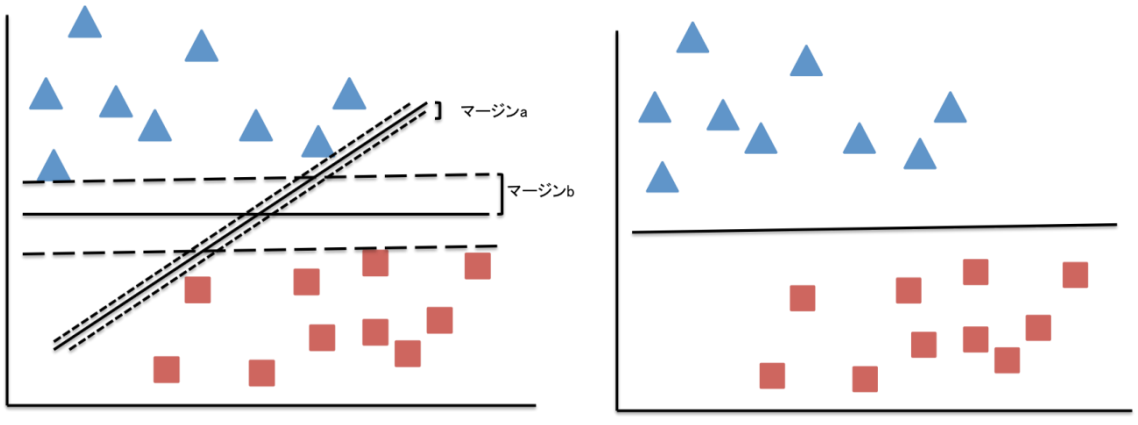


図 2.2　SVM：境界の決め方

上記図 2.2 左図のように，2 グループで最も距離の離れた箇所に線をひき，マージンが最大になる境界線を選ぶ．この場合，マージン b＞マージン a なので，マージンが大きい方のマージン b を選ぶことになる．そして図 2.2 の右図の境界線の真ん中に引いた線が 2 グループを識別する境界となる． SVM は境界を直接学習するために，学習データが少なくとも高い精度が得られやすい．

(3) *k*-NN

k 最近傍法とよばれる手法で，教師あり分類の 1 つである．似ている文書には同じカテゴリがつくという仮定に基づき分類を行う．
図 2.3 は，○の近傍にある N のカテゴリを調べ，数の多い方へ分類される例である．△と□は，あらかじめ教師データとして分類済みの文書と考え，そこへ，新しい文書○がどちらへ分類されるかを考える．
左は，N=5 の場合，○は数の多い△と同じカテゴリに分類されるが，右図 N=7 の場合，○は数の多い□と同じカテゴリに分類される[Tomiura 96]．

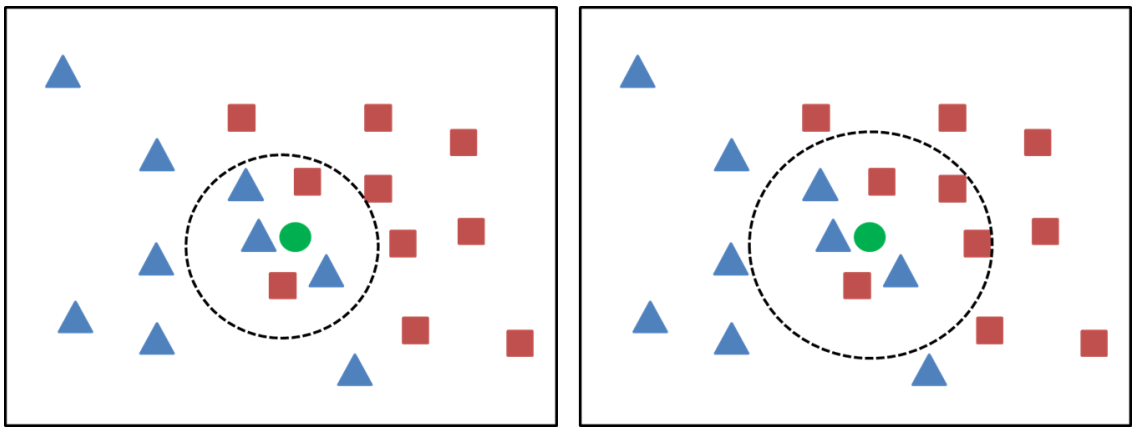


図 2.3　k-NN：カテゴリの考え方

k-NN は，教師データのノイズの影響を受けやすい．また，教師データの数

が少ない場合は，十分な精度が得にくい．ただし，逆に教師データの数が多い場合には高い分類精度が期待できるとされている．

2.1.3　コサイン類似度

コサイン類似度は文書の類似度を測る場合に利用される．その仕組みを述べる[Tan 06]．コサイン類似度は，ベクトル同士の成す角度の近さを表現するため，コサインが 1 に近ければ類似しており，0 に近ければ似ていないことになる．

コサイン類似度は，ベクトルの内積をそれぞれの大きさで割った式であらわされる（式 2.7）．

$$\cos\theta = \frac{X \cdot Y}{|\overrightarrow{X}| \cdot |\overrightarrow{Y}|} \qquad \text{（式　2.7）}$$

X,Y のそれぞれの文書中の要素をベクトル化する．たとえば，文書中に出現する名詞を要素とする場合（a,b,c,d はいずれも文書中に出現する名詞とする），

X= {a,b,b,c}　Y= {a,c,c,d}

のときは，

X= {1,2,1,0}　Y= {1,0,2,1}

のベクトルで表される．

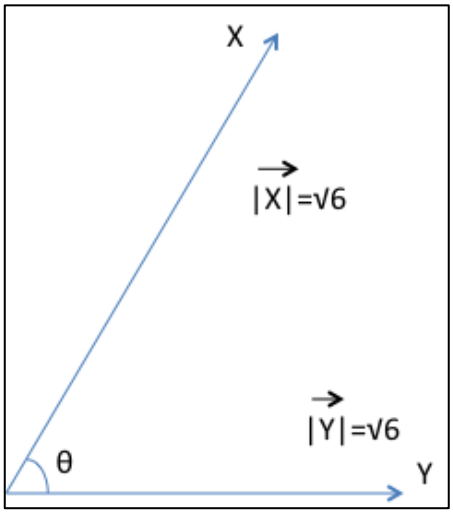


図 2.4　コサイン類似度の計算

図 2.4　によると，X と Y の内積は 3 なので，式 2.7 の計算により，コサインは 0.5　となる．

よって，計算上の文書 X と Y の類似度は，中間くらいの類似度をもつということになる．

2.1.4　性能評価

文書分類を含む自然言語処理の性能評価には, 多くの場合, 適合率(Precision)と再現率(Recall)が使われる[Kita 02]．適合率と再現率は，情報検索のための最も一般的な評価指標であり[Eguchi 04]，それぞれ，次のような意味を有している．まず，適合率は，検索質問に適合する文書だけを検索しているかを評価する尺度であり，検索された文書集合の中で，検索質問に適合する文書の割合を示している(式 2.8)．つまり，検索ノイズの少なさを示す尺度である．一方，再現率は，検索質問に適合する文書を漏れなく検索しているかを評価する尺度であり，検索対象となる文書集合の中の検索質問に適合する文書のうち，実際に検索された文書の割合を示している(式 2.9)．つまり，検索漏れの少なさを示す尺度である[Sakai 06]．適合率と再現率の概念を図 2.5 に示す.

また適合率と再現率は，一方の値があがると，一方の値が下がるというトレードオフの関係にあり，文書の再現性と適合性どちらを指標とするかによって評価も異なってくる．それえゆえ，2 値の調和平均をとる F 値（式 2.10）を参考にすることもある.

$$\text{適合率} = \frac{\text{検索された適合文書の数}R}{\text{検索結果の文書の数}N} \qquad \text{(式 2.8)}$$

$$\text{再現率} = \frac{\text{検索された適合文書の数}R}{\text{全対象文書中の正解文書の数}C} \qquad \text{(式 2.9)}$$

$$\text{F 値} = \frac{R}{\frac{1}{2}(N+C)} \qquad \text{(式 2.10)}$$

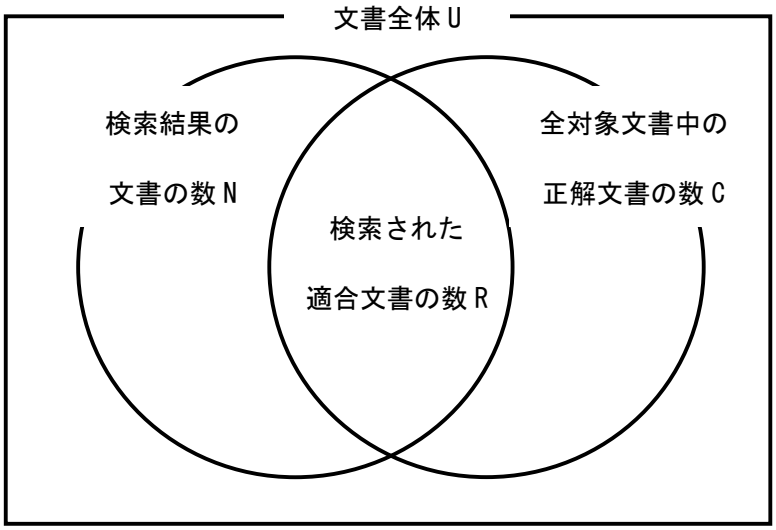


図 2.5　適合率と再現率

分類関数を利用し，文書を自動分類したとき，自動分類の結果と，人手で分類した結果とを比較して, 分類精度を計算する 2 つの方法を示す[Sebastiani 02].

・文書ごとの分割（document-pivoted categorization）
・カテゴリごとの分割（category-pivoted categorization）

下記図 2.6 によると，文書（D），あるいはカテゴリ（C）どちらを採用するか分類関数によって計算される.

分類関数（D またはC）		専門家による分類	
		T	F
自動分類	T	TP	FP
	F	FN	TN

図 2.6　分類関数と評価

TP, FP, FN, TN は，下記の意味を持つ．

TP(true positive)　検索結果がでてきて，かつ正解文書が得られた場合
FP(false positive)　検索結果がでてきたが，不正解の文書が得られた場合
FN(false negative)　出てほしい検索結果がでていない場合
TN(true negative)　出なくてよい結果が期待通り含まれていない場合

この場合，適合率(Precision)，再現率(Recall)の式は，次式に表される．

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (式\ 2.11)$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (式\ 2.12)$$

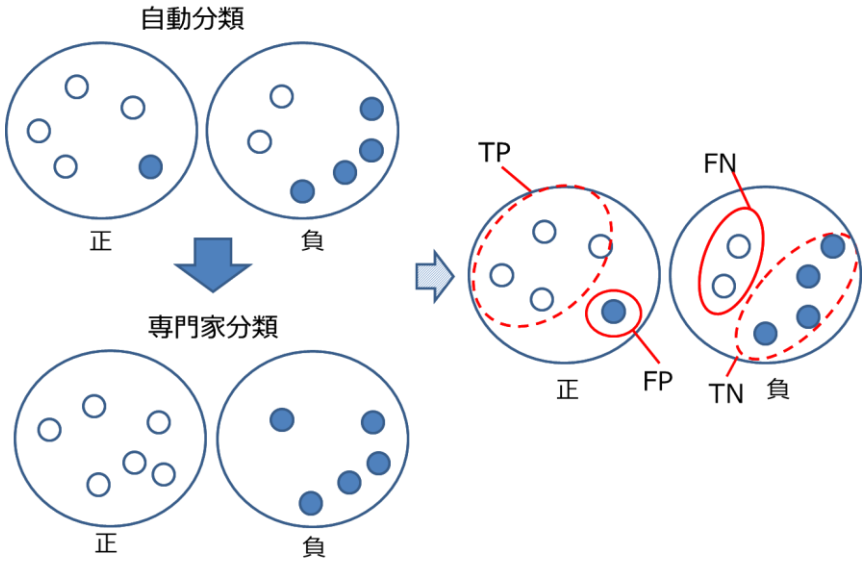


図 2.7　文書分類概念図

図 2.7 は，専門家の分類からみて自動分類を評価する場合の概念図である．
この図の例を考えると，

適合率　＝　4　/　（4　＋　1）＝　0.800

再現率　＝　4　/　（4　＋　2）＝　0.666

上記の結果となる．

このように，分類した結果を評価するシステムが確立されている．

2.1.5　検索システム

検索対象文書の主題を利用した検索や，検索対象文書の局所的な情報を考慮したパッセージ検索などがある[Hearst 93][Hearst 94][Iwayama 03].

(1)　全文検索

文書に含まれるすべての情報を対象に検索する方法であり，検索範囲が広いため検索漏れが少ないという利点があるが，条件を指定して絞り込みづらいため精度の面ではやや劣る．主な技術に索引型（インデックス型）がある．
この索引型に使われる索引を抽出する手法に，形態素解析や N-gram[Brown 92]が用いられている[Baba 98] [Kokubu 99].

(2)　概念検索

概念検索は，指定した文書と類似しているとされた文書群を類似性の高い順に抽出する技術である．

すでに，特許検索の分野では，全文検索の限界から概念検索を使った検索の応用例が検討されている[Muguruma 01] [Muguruma 03a][Muguruma 03b].
市川はビジネスモデル特許検索にもこの概念検索が有効だとしている[Ichikawa 01].

概念検索の概要を図 2.8 に示す.

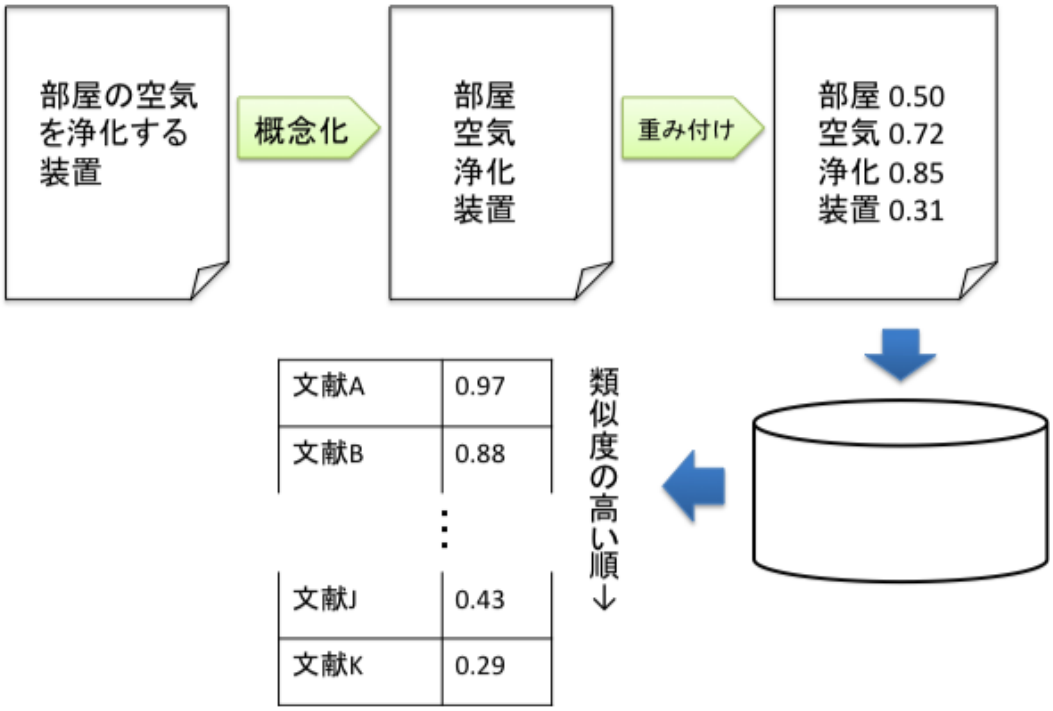


図 2.8　概念検索

文書を語群に概念化し，語群の語をそれぞれ重み付けする．そして，データベース中にある他の文書の中から，概念化した文書と近い類似度をもつ順に検索結果として返す[Kumamoto 99].

2.2　特許文書における文書解析と検索

特許文書の解析について,1990 年代あたりから現在に至るまで多くの研究がなされている[Yamaoka 98][Sakachi 10][Okumura 12].

その目的は，特許先行技術調査に利用することである．特許文書の中でも特許請求の範囲として書かれている請求項について解析されることが多い．その理由として，請求項が，発明の権利範囲を決定する重要な箇所であり，必ず理解しなければならないポイントであるのに対して，非常にわかりづらく書かれているからである．

本節では，主な研究の概要を記し，一般的な文書と特許文書の解析と検索の違いを明確にする．

2.2.1　特許請求の範囲の構造

日本では，特許請求の範囲として知られる請求項の書き方として，特段の規定はないが，一般的に書かれる書き方として知られているものがある．以下，主な例をあげる．

(1) ジェプソン形式

「～において」「～において」の前後をわけ，前文を先行研究と共通する既知の部分，後文を出願における特徴部分と書きわける形式.
この記載形式は非常に多く目にすることができる．

記載例：
A をもつ XX 装置において，B をする手段と，C をする手段と，D をする手段を備えることを特徴とする XX 装置.

上記例の場合は，A をもつ XX 装置が既知の部分となり，B, C, D を備えることで XX 装置が新規性をもつことを主張している．

(2) 順次列挙形式

時系列的に構成要素を順次説明していく形式.

記載例：
A に B を設けるとともに，この B の端部に C を設け，この C に D を取り付けて構成した XX 装置.

(3) 構成要素列挙形式（一項記載）

一項ずつ，構成要素を並べて書く形式

記載例：

A と，

この A を設けた B と，

この B の端部に設けた C と，

この C に取り付けた D と，

を備えた XX 装置.

(4) 構成要素列挙形式（箇条書）

(3)を箇条書きのような形成で書く

記載例：

以下の構成を備えた XX 装置.

(1)A

(2)前記 A を設けた B と，

(3)前記 B の端部に設けた C と，

(4)前記 C に取り付けた D と，

(5) 変形構成要素列挙形式

さきに構成要素を並べて書き，あとから各要素の説明を加えて書く形式.

記載例：

A と，

B と，

C とを備え，

前記 A は・・，

前記 B は・・,
前記 C は・・,
であることを特徴とする XX 装置.

2.2.2　特許文書分類

大量にある特許文書に対して，機械学習を利用し分類を行う手法が多く見られる[Fall 03][Chu 08][Wu 10]．以下に，主な分類手法を示す．

特許の分類体系は，階層化している．そのため，まずは上位分類から絞り込んでいき，細かい階層部分には，階層レベルごとの分類器を構築し，下位分類を行っていく方法がとられる[Koller 97] [Yoshida 05][Koyano 07][Usui 13].

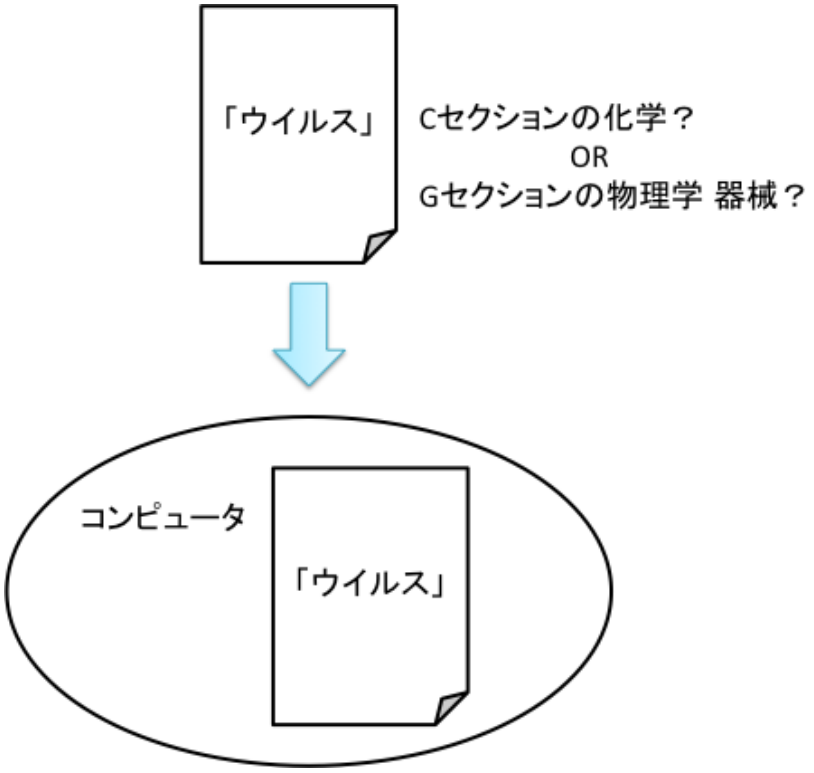


図 2.9　階層利用の利点

図 2.9 のような，インフルエンザウイルスのウイルスなのか，コンピュータウイルスのウイルスなのか分類のわからない語についても，その語をもつ文書のカテゴリ（コンピュータ）がわかっていれば，分類の第一段階として，この階層を利用できる．

2.1.2 に記した，文書分類(1)のナイーブベイズ法を利用し，単語 Wi がカテゴリ C に出現する確率 P(W | C) を決める際に，C の上位カテゴリ C_{up} を用いて，このとき$P(W|Cup)$(式 2.13) との線形和でスムージングを行い，確率をだす shrinkage 法を McCallum らが提案した[McCallum 98b].

$$P(W|Cup) = \frac{単語Wiの数}{カテゴリCupの語彙の数} \qquad (式\ 2.13)$$

Cai らは, SVM に階層を利用し, カテゴリ C をもつ文書は上位カテゴリ Cup ももつものとして分類学習させ，その結果，階層をとりいれる手法のほうが階層をとりいれない手法よりも高い分類精度が得られるとした[Cai 04].

2.2.3　文書間類似

文書間の類似は，文書中に出現する単語と単語の類似度によって決まってくる．しかし，この方法だと，文書中の単語と単語の類似を総当たり計算する必要があり，計算も複雑になってしまう．[Oinuma 13]は，あらかじめ，一般的な単語のセット同士の類似度を計算しておき，それを対象文書に含まれる単語と照らし合わせ概念ベクトルを生成し，文書間の概念ベクトル同士で類似度を算出することで，計算の煩雑さをカバーした．ここでは，あらかじめ作成した単語セットを新聞記事により作成したため，特許文書で有効かどうかは示されていない．[Yanagimoto 13]は，ニューラルネットワークを利用した文書類似度の推定を行った．大量の文書を学習し，入力された単語の出現頻度で定義された特徴ベクトルを低次元の特徴ベクトルで表現し，最終的に得られた特徴ベクトルを用いて文書間の類似度を計算する方法で, 文書間類似度の改善を行った．しかし，学習という概念のあるニューラルネットワークを応用する手法では，学習データに忠実なあまりテストするデータでは予測できなくなるという過学習の問題がある．

特許文書の文書間類似について，[Suzuki 13]は，同義語や言い換えなどのシソーラス情報を用いず，対象となる文書中のキーワードの出現頻度パターンはコレスポンデンス分析を用いて抽出し，この結果をもとにした検索語の合成スコアを求め，この合成スコアとキーワード頻度パターンの類似性で文書検索する手法を提案した．そして，従来のキーワード検索との定量的な比較を行い，F 値が高く求められた．ただし，この検証は，対象特許文書が「要約」部分に限られており，発明の核となる請求項についての検証ではない．また特許文書中の不要語については削除しないで行ったため，不要語の検証への影響は不明である．また，[Uneda 12]は，「請求項」と「詳細説明」の自動対応付けを提案する際に，文書ペアの類似度から対応付けする手法を提案している．4 つの文書間の類似性尺度と 3 種類の機械学習の組み合わせにより，類似する請求項と詳細説明を抽出した．組合せ手法では，SVM を利用し，かつ文書間の形態素数の比と手がかり句となる語の頻度を用いた場合が，F 値が一番高くなった．

ただし，2 つの文書が，表層的に類似ではないが，内容的に類似であるものについて，システムでは検出できなかった．逆に，表層的に類似であっても，内容的に非類似であるものを類似としてしまう場合もあり，表層的類似だけでなく，意味的類似についても検討する必要がある．

2.2.4　特許文書分析

特許文書の中に書かれている発明の権利範囲を決める請求項は，特許文書の中で非常に重要な箇所である．よって，この請求項を正しく解釈することが，特許文書解析の肝となる．

しかしながら，この請求項は長い名詞句で書かれており，途中で句点はい．そして，名詞や名詞が連なってできている複合名詞の前に「あの」「この」など修飾が曖昧になるおそれのある指示代名詞は使われず，かわりに「前記」「該」「当該」などが使われている．そして，発明の新規性を主張するために，出願者が意図的に使う，未知な複合名詞や新語の存在．これらが複雑に絡み合って，請求項の読解を難しいものとしている．

図 2.10 に実際の請求項の記載例を示す．

【特許請求の範囲】
【請求項１】
生体認証用の生体認証情報を認証の対象となる対象者の体から取得し、前記体の位置を検出する生体認証情報取得部と、
前記取得した生体認証情報と照合する照合用の情報を、前記対象者を個々に識別する個人識別情報に対応付けて複数人分記憶している記憶部と、
前記個人識別情報を入力する入力部と、
前記入力部を介して入力される個人識別情報に対応する照合用の情報と、前記生体認証情報取得部により取得された生体認証情報とを照合して個人認証を行う認証部とを有する生体認証装置であって、
前記照合用の情報を作成する際に使用した機器の使用条件および前記対象者の体に関する情報の少なくともーつを設定データとして、前記個人識別情報又は前記照合用の情報に対応付けて記憶している設定データ記憶部と、
前記認証時に入力される個人識別情報または当該個人識別情報により特定される前記照合用の情報に基づいて前記設定データを読み出し、当該読み出した設定情報の少なくともー部を前記生体認証情報取得部に設定して前記生体認証情報を取得させる設定部とをさらに有し、
前記認証部は、前記生体認証情報取得部より取得された生体認証情報から生体の位置情報を取得する位置情報検出部を有し、
前記設定データ記憶部は、前記設定データの１つとして、前記照合用の情報を取得したときにおける前記生体の位置情報を有し、
前記認証部は、前記照合用の情報を取得したときにおける前記生体の位置情報と、認証時における前記生体の位置情報とを比較して、前記生体の位置の変化量を算出し、前記生体の位置の変化量が所定のしきい値以下の場合には、前記生体の位置情報以外の設定データを前記生体認証情報取得部に設定した状態で取得された生体認証情報を用いて個人認証を行うことを特徴とする生体認証装置。

図 2.10　公開特許公報 2012-073905 請求項 1

この例のように，句点「。」は文の最後のみに出現し，名詞，複合名詞の前に置かれる「前記」「当該」が 30 回以上出現している．それに加えて「生体認証情報取得部」など長い名詞連なりである複合名詞も出現している．

ここで請求項を難読なものにしている主な理由を整理すると，2 つの原因に絞られる．

・名詞句が長いため，文の前後の修飾関係がわかりにくい
・未知なる複合名詞の存在により，その語が何を示しているかがわかりにくい

1 つ目の問題である，請求項の一文が長すぎるために起こる，読みづらさに対して，可読性を向上するための研究が行われている．
通常の文の係り受け解析で使われるような解析器 KNP[Kawahara 06]を用いると，修飾部が長すぎてエラーを起こしてしまう．そのため，句読点で区切りのある短文を含む文書を係り受け解析する方法を特許文書に適用することは難しい．橋本らは，特許文を既存の解析器に適応させるために，各述語の各要素（ガ各やヲ各など）に対してどのような語が用いられているかという格フレーム法を用いた[Hashimoto 08]．

請求項を読解しやすくするために，新森らは，請求項の特徴的な記述特性を生かして，請求項の中にある手がかり句（「において,」「であって,」「を備えた」「を特徴とした」など）を用いて請求項文を構造解析した[Shinmori 04]．その例を図 2.11 に示す．

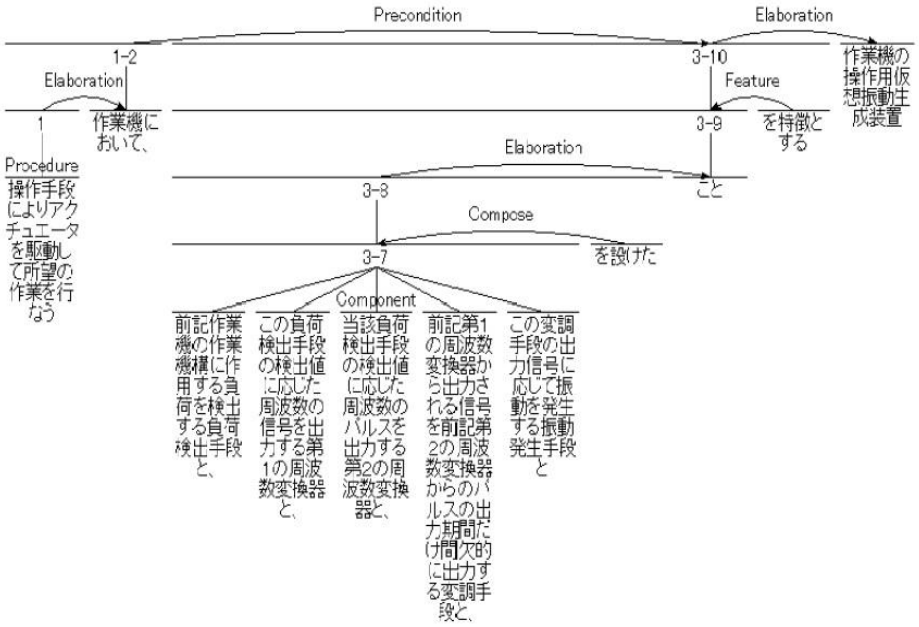


図 2.11　新森らの研究 手がかり句による請求項構造解析の例

新森らは，この手がかり句をもとにした文法規則を 57 個作り，それら規則はそれぞれ独立しているため，図 2.11 のように曖昧性なく，読みやすく構造ごとに区切る結果を得ることができている．

特にジェプソン形式における解析については，高い精度で正しい解析が行われたという結果が得られたが，これらはジェプソン形式に出現する手がかり句の「において,」「であって,」が重要な役割をもっているためとしている．

また，同じく，請求項の理解支援のために，小西らは[Konishi 06]，パターンマッチングを利用し，段階的に請求項を属性ごとに整理し，最終的に意味を持つ属性にまとめ，それをリスト表示することで請求項が見やすくなるという手法を提案した．

2 つ目の問題である未知語，新語とされる複合名詞の意味を判別する問題について, 藤井らは, それまで, Web から言葉や事柄に関する説明情報を抽出し，さらに複数の説明情報を組織化することで，百科事典的なコンテンツを自動構築する研究をしてきた[Fujii 02]ことを踏まえ，特許情報から用語辞典的なコンテンツを自動構築し, 専門用語の知識として活用するシステムを提案した[Fujii 08]．この目的は，Web では抽出できないような請求項中の複合名詞の説明情報を特許情報から収集することである．このような複合名詞は専門用語として使われるため，特許情報以外の情報からその複合名詞に対する説明情報が得られにくい．

複合名詞に限らず，新語の同定について，橋本らは，特許文書に特化した辞書を作成するため，作成した規則に従い形態素の候補語を選定し，従来の形態素解析辞書と k-NN 法によって新語の品詞の同定を行った[Hashimoto 12]．

また，難波らは，学術論文に出現する語を特許文書に出現する語と紐づけて変換する手法を提案した[Nanba 09]．例えば，学術論文において，「ワードプロセッサ」を入力すると「文書編集装置」や「文書作成支援装置」に変換する技術である．これにより，通常の文書中には見られにくい語である「文書作成支援装置」が「ワードプロセッサ」のことと理解することができる．

ここでは，引用手法[Nanba 05]，シソーラス手法[Nanba 07]，間瀬らの手法

[Mase 07]を組み合わせて方法がとられている．

小西らは，請求項に出現する語を説明している部分を発明の詳細説明の部分から探し出して，その中で使われている語を検索クエリに加えることで抽象的な表現で書かれている請求項中の語の意味を補う[Konishi 04]．これにより請求項と請求項の比較をする際に生じる，用語不一致のための検索漏れが軽減する可能性をもつことができた．小西の手法を図 2.12 に示す．

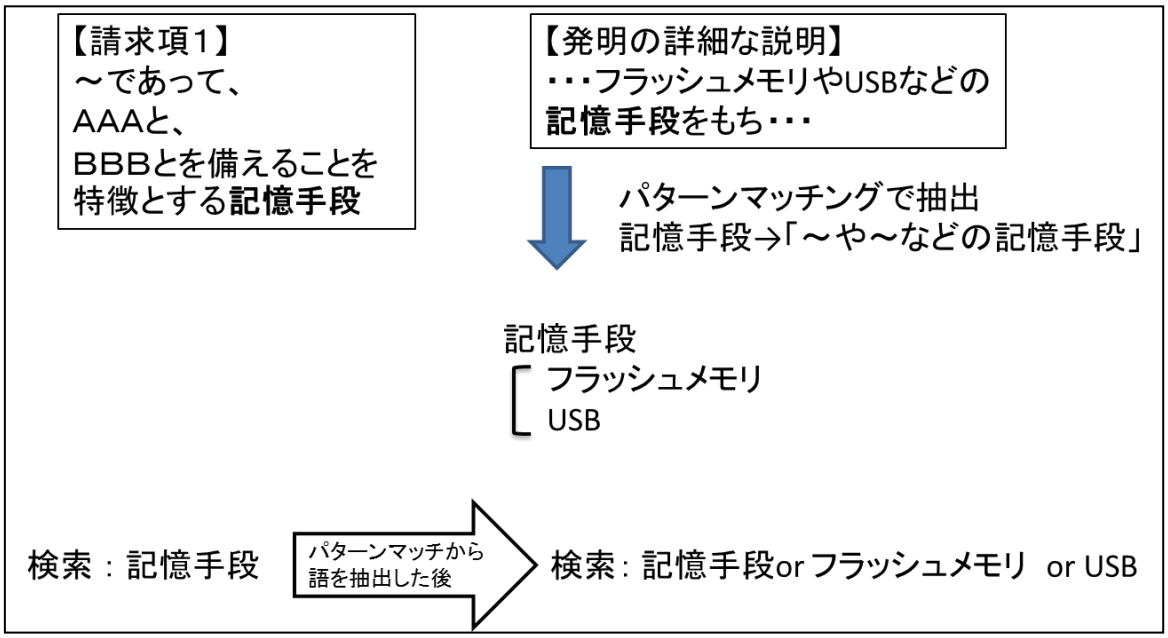


図 2.12　小西の手法

2.2.5　複合名詞の特徴

2.2.3 の特許文書分析の中で取り上げた請求項内に出現する複合名詞について，複合名詞そのものについての意味構成，または，複合名詞を文書の中から抽出する手法など，複合名詞にまつわる研究をあげておく．

(1) 複合名詞の重要度に関する研究

日本語の中には，多くの複合名詞が出現するが，そのすべてを網羅し，辞書にするのは容易ではない．複合名詞を構成している語は，単に共起しているというより係り受けのような密接な関係をもっている．このような情報を検索に有効に生かすために，複合名詞の重要度を算出する方法を提案した[Yamada 96]．この目的は，文書検索で文書の重要度を考慮するときに，複合名詞は複合名詞のまま重要度を考えれば，より文書検索の精度が上がると考えられたことによるものである．

(2) 複合名詞抽出に関する研究

村上らは，テキストマイニングの世界で，単名詞が 1 つの分析単位とする手法で使われる形態素解析では，分析単位によって結果が大きく左右されるため，単名詞だけではなく，単名詞から構成される複合名詞など拡張された概念を利用することにより正確な分析が可能になるとしている[Murakami 06]．そして複合名詞を構成する単名詞と比較し，時系列で出現するポイントの相関を確認し，相関があった場合は，複合名詞がその単語に対する時系列発生的であるとした．

中川らは，文書から複合名詞を抽出する研究をおこなっている[Nakagawa 03][Yumoto 01]．中川らの手法を利用した応用研究を特許文書に適用した例もある[Takemori 10][Kato 06]．

(3) 複合名詞の構成に関する研究

2 名詞で構成される複合名詞について，国語的な生成文法については以前より研究されている[Okutsu 75][Saito 92]．小林らは，複合名詞を構成する名詞を分割し，意味をもつ可能性のある名詞を意味分類辞書により分類し，コーパスに出現した共起頻度をもとにした優先度を計算することにより複合名詞の解析を行った [Kobayashi 96]．

また，竹内らは，語彙概念構造を利用して係り受け関係を解析した．図 2.13 の例のように一方の名詞が他方の名詞にどのように働くか解析した[Takeuchi 02]．

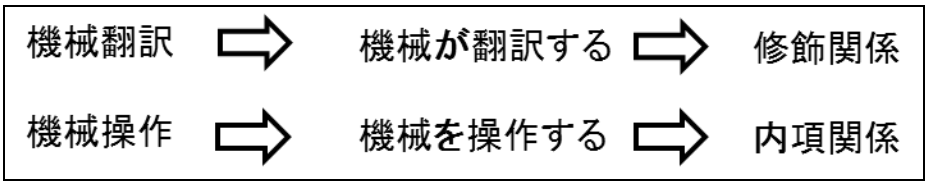


図 2.13　複合名詞内の語関係

特許文書中に出現する複合名詞の解析では，内山らが，2 名詞から成る複合名詞について，各名詞の意味情報と文法から意味関係解析規則を作成した [Uchiyama 06a][Uchiyama 06b]．

2.2.6　特許検索と活用

一般的に簡単に特許検索をする場合，特許庁データベースが用いられる[J-PlatPat 15]．検索の手法も，素人ではなく，専門家が長年の経験やコツやノウハウを駆使して特許検索を行うことが主流である[Sakai 07]．しかしながら，この方法では，検索する人によって結果が異なってしまうことにもなる．

特許検索システム自体については，難波らが，特許データベースと特許検索履歴からシソーラスを自動的に構築する手法を提案している[Nanba 11]．

また，間瀬らは，ジェプソン形式で書かれた請求項を「において」で既知となる前提部分と本題が書かれている本質部にわけ，本質部を使って検索を行い，対象文書を絞り込む方法を提案した[Mase 05]．

検索システムの応用例として，自社技術の障害となるような他社特許を無効化するための公知資料を探す調査を無効特許資料調査というが，そのための無効特許検索システム[Takagi 05]などもある．

また，佐藤らは，特許間の引用を，出願人ごと，もしくは引用数ごとなどで分類し，分類ごとに特許の重要度を算出する方法で重要特許を検出する手法を提案した[Sato 06]．

検索した特許情報を活用するという視点では，情報を可視化し，出願技術傾向を探り，技術マップを作製するなどの研究が行われている[Kiriyama 06][Ando 07][Kunishi 11][Sato 14]．　安藤[Ando 09]は，エンドユーザーが利用できる方法で類似度による公報の解析の可視化と発明者と特徴語のネットワーク分析を行った．

一貫して考えると，発明や研究内容を熟知している出願者や発明者でもあるエンドユーザーが，外注頼りではなく，自ら検索から活用までできることが望ましい．また，そのためのスキル教育の必要性も各会社，産学の場でも提案されている[Ishida 04][Deguchi 06][Okamoto 06]．

2.3　結言

特許文書と自然言語処理との関係において，一般的な文書と特許文書，主に請求項を比較し，

・請求項の構造
・請求項中の複合名詞

上記 2 つの問題が，類似文書検索するうえで必要となる請求項理解の妨げとなっている．

類似文書を検索する場合，構造を理解する側面からのアプローチと複合名詞を形態素に代わって利用するアプローチがあることは理解できる．文書の構造については，新森，小西らの研究により，読みやすさの部分でかなり成果を上げている．

複合名詞については，その重要性は認識され，複合名詞のもつ意味を判定する手法も研究されている．ただし，4 文字を超えるような複合名詞を対象にすると，複合名詞を構成する単名詞の分割箇所によって 2 通りの意味を生じたりする．

そして，特許文書のみならず一般文書においても，複合名詞を利用することの有用性は確実には示されてはいない[Sebastiani 02]．

多くの先行研究において，確実な精度をもった検索が行えていない理由は，特許文書に合った辞書が整備されていないことに起因する．特殊な用語が多用されている特許文書について，辞書の整備は不可欠である．この辞書について検討していかなければならない．辞書整備を含めた課題を次章以降で述べる．

そこで，さらに効率的な特許文書検索を行ううえで，解決しなければならない問題を以下にまとめた．

(1) 複合名詞を使う類似検索の有用性
(2) 複合名詞の類似データの作成
(3) 複合名詞を使った類似文書検索の結果

上記(1)の解決法については,「第 4 章 複合名詞を利用した文書類似検索の可能性」にて検討する．上記(2)の解決法については,「第 5 章　複合名詞の類似判定方法」にて検討する．上記(3)の解決法については,「第 6 章　複合名詞を利用した文書類似判定」にて検討する．

第３章

特許制度および特許分類

本章では，特許出願と出願者が査定を受けるまでの過程，特許文書解析に必要となる分類，そして特許文書解析の必要性について述べる．

ここでは，特許制度の流れを理解し，先行技術調査の意義を確認することで本研究の対象を明確にする．

また，本研究では，特許文書中に出現する類似複合名詞を拒絶理由通知書から抽出することで同定し，類似特許検索への有用性を示すことを目標としている．そのために拒絶理由通知書について説明し，必要な複合名詞のデータをどのように抽出できるかについて示す．

3.1　特許出願のながれ

本節では，特許出願の概略，特許出願の流れを示す[JPO 15].

特許審査は出願したのち，自動的に行われるのではなく，出願後に出願者が審査請求することではじめて開始される．そして，最終的には出願した特許に対して審査官から「特許査定」あるいは「拒絶査定」を受ける．

通常，出願された特許出願は 18 ヶ月後に公開される．その際に公開される文書が「公開特許公報」である．出願者は，この出願から 3 年のうちに実際に審査請求をするかどうかを決定する．審査請求しない場合は，その出願が取り下げられたものと見なされる．請求をした場合は，書式などの不備がないかなどを確認する方式審査を経たのち，不備がなく却下の理由がない場合，次なる実体審査にうつる．実態審査ののち，最終査定を受けることとなる．

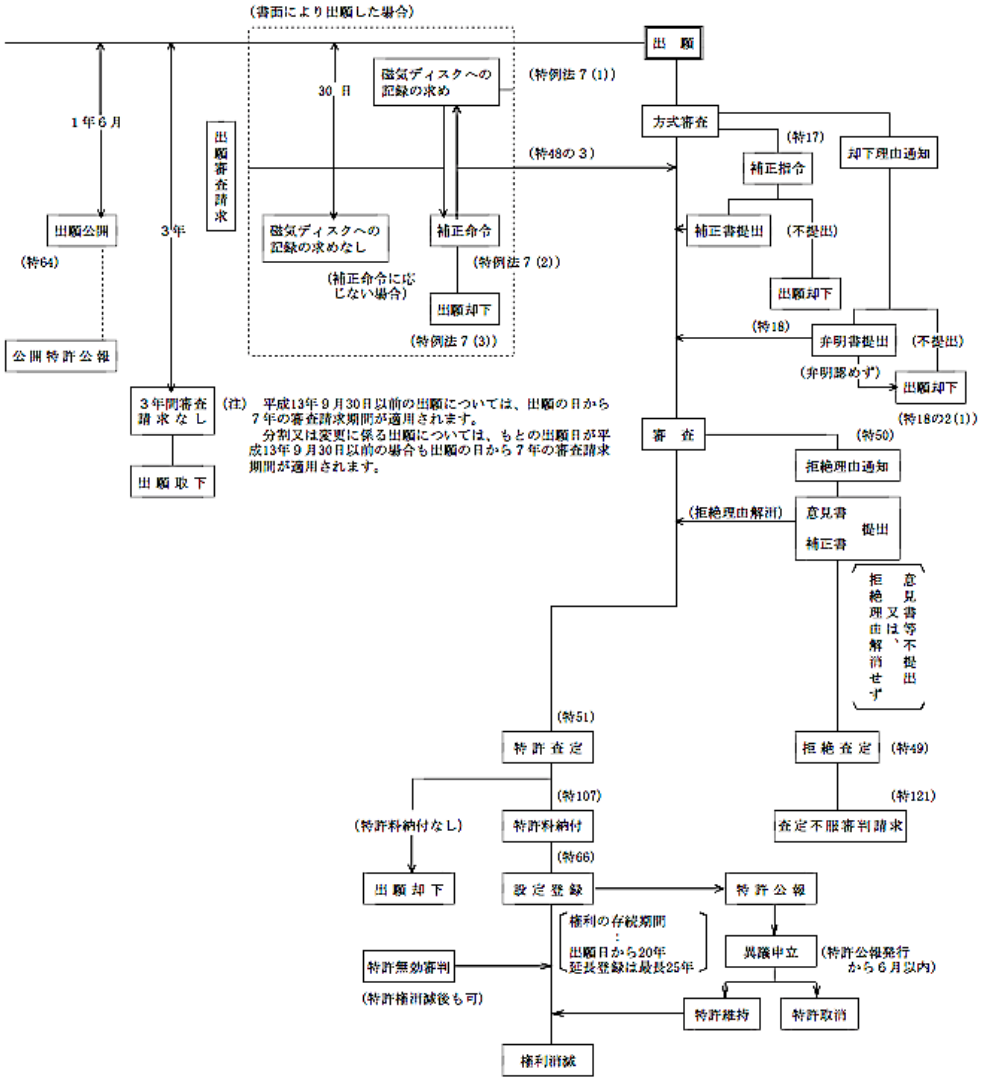


図 3.1　特許審査のながれ

出願された特許を審査官が拒絶するときは, 特許法 49 条[Tokkyocho 12]に従い，拒絶査定を行わなければならない．逆に，拒絶に値する理由がないときには特許法 51 条により，特許査定を行わなければならない.

特許を取得するための特許要件を，図 3.2 に示す.

各特許要件のいずれかを満たさない場合は拒絶されることとなる.

例えば，発明の成立性で新規性をみるときには，その発明が新しいかどうか，特許法 29 条第 1 項を満たすのであれば，新規性における特許要件を満たすこ

ととなる．ただし，29 条の各号に該当する場合は，特許要件を満たすことができないので，拒絶されることとなる．特許法 29 条第 1 項は下記のとおりである．

第 29 条　産業上利用することができる発明をした者は，次に掲げる発明を除き，その発明について特許を受けることができる．

一　特許出願前に日本国内又は外国において公然知られた発明

二　特許出願前に日本国内又は外国において公然実施をされた発明

三　特許出願前に日本国内又は外国において，頒布された刊行物に記載された発明又は電気通信回線を通じて公衆に利用可能となった発明

図 3.2　特許要件図［Yoshifuji 97］

審査官が拒絶しようとする旨を出願人に通知する場合，出願人に対していきなり拒絶査定を行うのは，酷なことであるから，特許法 50 条により拒絶の理由を通知し，一定の期間を設け出願人が意見する機会が設けられている．このときに通知される文書を「拒絶理由通知書」という．

3.2　国際特許分類

本節では，特許分類に必要となる国際特許分類について述べる．

特許分野を分類するため，国際的な分類基準として世界共通の国際特許分類番号（IPC）が定められている[WIPO 14]．

IPC は，特許の分野に適当であると認められる全知識体系を８つのセクションに分けて表現している．セクションは，IPC 階層の中で最も高い階層にあるものである．

(a) セクション記号 -各セクションは大文字 A から H のうちの１つで表示される．
(b) セクションタイトル -セクションタイトルはそのセクションの内容をごく大まかに指示するも
のとしている．

８つのセクションは次のとおりタイトルが付けられている．

A 生活必需品
B 処理操作；運輸
C 化学；冶金
D 繊維；紙
E 固定構造物
F 機械工学；照明；加熱；武器；爆破
G 物理学
H 電気

各セクションは「クラス」に細分化され，クラスは，１つ以上の「サブクラス」

を含む．そして，各サブクラスは「グループ」に細展開項目に展開される．グループは，「メイングループ」と，その下位層にあたる「サブグループ」とで構成される．

A01B33/00 の場合，

A　　セクション
01　　クラス
B　　サブクラス
33　　グループ
00　　メイングループ（もしくはサブグループ）

図 3.3 に，　A01B33/00 イメージを示す．

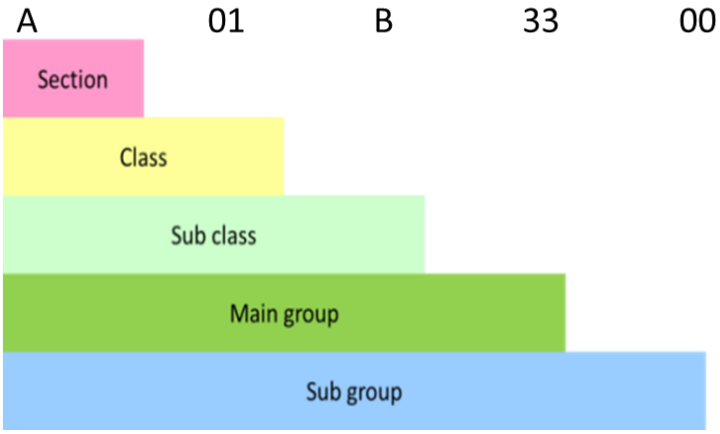


図 3.3　IPC 分類階層

A01B は，農業 林業における土作業：農業機械 器具の部品，細部 附属具一般の項目であり，グループ 33/00 その中のメイングループの駆動回転具を有する耕うん用作業機の分野を示す．

このように，分類は非常に細分化されており，文書分類や検索の際にも，この階層構造を用いられる．

また，日本独自の分類である F ターム，FI などの分類も存在するが，ここでは，割愛する[Muto 00].

出願者，および特許審査官は，対象文書がどの分野のものであるか IPC をもとに分類する．分類番号の付与は 1 つとは限らないが，3 つ〜5 つほど付与されているものが多い.

3.3　先行技術調査の意義

本節では，特許出願にあたり，出願者が行う先行技術調査の意義について述べる．

先行技術調査は，出願時だけではなく，出願前に特許出願しようとする者が，その発明を研究する前段階に行うことも先行技術調査に含まれる．なぜ，先行技術調査を行うのか，現在の特許出願とその結果，そして先行技術調査が不足するとどのような弊害が生じるのかということから，その意義を考える．

(1) 拒絶査定

図 3.4 は，2006 年から 2010 年までの特許出願に対する最終処分実績の推移表である．

特許査定の数は，5 年間でわずかながらに上がっているとはいえ，まだまだ拒絶査定の数も半数近くに上っている状態である．

【最終処分実績の推移】

実　績	2006年	2007年	2008年	2009年	2010年	前年比
特許査定件数	129,071	146,383	159,961	178,227	205,652	115%
拒絶査定件数	129,400	147,678	154,163	171,396	164,639	96%
(うち応答なし拒絶査定件数)	68,879	78,246	85,443	105,004	100,951	96%
FA後取下げ・放棄件数	7,915	5,567	4,779	5,169	4,600	89%
特許査定率	48.5%	48.9%	50.2%	50.2%	54.9%	−
拒絶査定率	51.5%	51.1%	49.8%	49.8%	45.1%	−

（備考）・応答なし拒絶査定件数とは、審査官の拒絶理由通知に対し、何ら応答されず拒絶査定された件数。
・FA後取下げ・放棄件数とは、一次審査着手後に出願の取下げ・放棄が行われた件数。
・特許査定率＝特許査定件数／（特許査定件数＋拒絶査定件数＋FA後取下げ・放棄件数）
・拒絶査定率＝（拒絶査定件数＋FA後取下げ・放棄件数）／（特許査定件数＋拒絶査定件数＋FA後取下げ・放棄件数）
（資料）特許庁作成

図 3.4　出願最終処分実績の推移(特許行政年次報告書 2011 年版　特許庁)

拒絶査定を受けると，その出願に対して特許権を得ることはできない．
拒絶査定を受けることにより，出願者がそれまでに費やしたであろう，人件費，時間，研究にかかるコストなどが無駄になってしまう．

また，特許権を取れるであろうことを想定した動きによっては，差し止め請求や特許権侵害訴訟をおこされる恐れもある．

(2) 損害賠償請求の例

自身が特許権を持っていない場合で，他者が特許権をもっている発明と類似するような発明を行っていた場合，それを知らずして行っていた場合でも，特許権をもっている他者から，特許侵害訴訟を起こされる場合がある．

図 3.5 に日本とアメリカにおける特許侵害訴訟における損害賠償額の例を示す．

日本

	原告	被告	技術分野	損害賠償額
1	アルゼ㈱	サミー㈱	機械(プログラム)	74億1800万円
2	スミス・クライン・アンド・フレンチ・ラボラトリーズ・リミテッド/スミスクライン・ビーチャム製薬株式会社	藤本製薬㈱	化学	30億5900万円
3	㈱観和製作所	フルタ電機㈱	機械	12億4400万円
4	アルゼ㈱	㈱ネット	機械(プログラム)	9億8800万円
5	マニー㈱	㈱秋山製作所	機械	7億1500万円
6	石川島播磨重工業㈱	㈱日立製作所	機械	4億3300万円
7	㈱観和製作所	フルタ電機㈱	機械	3億8200万円
8	三水㈱	リンテック㈱	化学	3億8700万円
9	テルモ㈱	バイエルメディカル㈱	機械	3億3800万円
10	三菱電機㈱	ダイニチ工業㈱	機械	2億8400万円

アメリカ

	原告	被告	技術分野	賠償額
1	セントコア	アボット	バイオ	1338億円
2	ルーセント	マイクロソフト	情報	1200億円
3	ミラー・ワールド	アップル	情報	500億円
4	サブラン	ボストン・サイエンス	バイオ	345億円
5	ユニロック	マイクロソフト	情報	310億円
6	ルーセント	マイクロソフト	情報	284億円
7	ハイニックス	ランバス	情報	245億円
8	メトロニック	ボストン・サイエンス	バイオ	200億円
9	デビュー	メドトロニック	バイオ	181億円
10	141	マイクロソフト	情報	160億円

図 3.5　日米損害賠償額比較例[Baba 15]

日本では，高額な損害賠償額に発展する例はまだ少ないが，今後，パテントトロール問題が横行する米国のように巨額な請求を起こされる可能性がないわけではない[Hiratsuka 09][Hiratsuka 13].

先行技術調査が十分でないと，出願に対して簡単に拒絶査定を受けてしまう結果となる，あるいは，損害賠償請求などの不測の事態も起こり得る.

このような例より，実際に特許出願しようとする前，あるいは出願を見据えての研究を開始しようとする前には，十分な調査をする必要がある.

3.4　先行技術調査の目的と方法

本節では, なぜ, 特許出願の際に先行技術調査をする必要があるのかを示し, そしてその方法を具体的に示すことで，調査の目的と問題点を明確にする.

3.4.1　先行技術調査の目的

出願者の立場から，先行技術調査をする目的，なぜ先行技術調査が必要なのかを述べる.

⑴特許権獲得の可能性を判断するため

出願が拒絶される理由のほとんどは先行技術の存在による新規性・進歩性の欠如である．すでに同じような発明についての出願がされている場合は特許出願が無駄になってしまう．そのため，出願の無駄をなくすことは出願費用の無駄をなくすだけでなく研究開発に要する金銭的・時間的なコストの無駄もなくすことを意味する.

また，先行技術を把握することで新たな改良発明を生み出す可能性もでてくる.

⑵特許文書作成の参考資料として利用するため

出願をする場合に同じ分野の発明に関する他の特許文書がどのように書かれているかを参考にする場合があり，どのように記載すれば特許査定を受けて特許権を得られるか，またどのように記載して拒絶査定となっているかなどの参考にもなるために他者特許を調査する.

(3)権利侵害とならないかを判断するため

仮に特許権がとれない場合でも権利侵害となっていなければ発明は実施できる.

また逆に，特許権がとれる場合であっても，先行する特許権等の利用発明である場合は自由に発明を実施することができない.

従って，出願しようとする発明の実施の可能性を判断するために，出願するしないにかかわらず，調査をすることが重要となる.

3.4.2　先行技術調査の方法

出願者の立場から，先行技術調査をするための具体的な方法について簡単に説明する．

(1)インターネットを利用する

インターネットを使用できる環境にあるならば，特許庁の特許電子図書館(IPDL は 2015 年 3 月で終了，現在は J-PlatPat としてサービス提供)で検索をする．

欠点としては平成 5 年以前の公報の内容をチェックすることができない．公報の番号がわかっているならば，以前の公報はヨーロッパ特許庁の esp@cenet でもみることができる．

そのほかにも無料のものとしてアメリカの特許庁のホームページで米国出願等についての検索ができる．有料のサービスではあるが，パトリス Web や NRI サイバーパテントデスクなども使うことができる．図 3.6 は，WIPO の検索ページであり，こちらも無料で利用することができる．

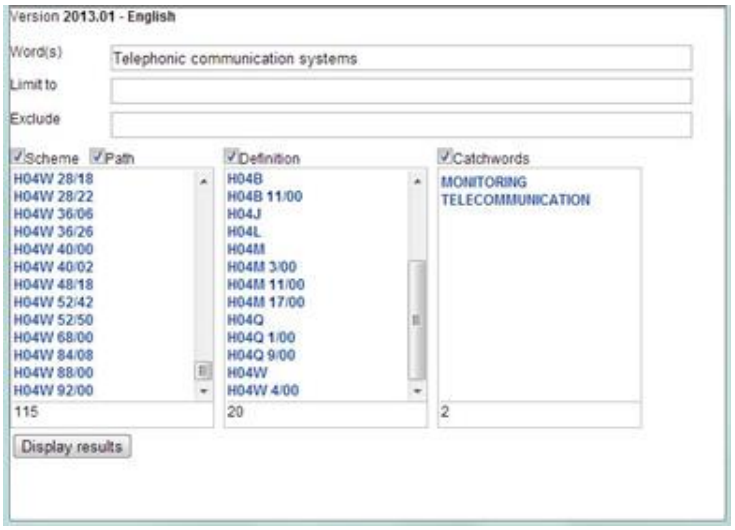


図 3.6　WIPO 検索ページ

(2)公報を直接調査

特許庁や全国の知的所有権センターもしくは発明協会へ行くと発行された公報を見ることができる．公報は最新のもの以外は IPC により分類されているので各年度分を検索できる．
また，特許庁などでは国内のみならず，各国の特許公報を見ることができる．インターネットで見ることができない昔の公報を確認したい場合などは有効な調査方法といえる．

(3)外部調査機関に依頼する

時間がない場合や調査にコストがかかっても構わない場合は，専門の外部調査機関に依頼することもできる．特許庁ホームページの関連ホームページリンクの特許情報提供事業者リスト集で調査機関を検索することができる．特許事務所に直接依頼する方法もある．

先行技術調査はいずれも人手を介す方法がとられる．そのため調査の結果が，調査する人間の主観，あるいは経験と知識に左右されてしまう．

この問題を解決するために，人の経験や知識を利用する方法に頼らずに調査する方法が必要といえる．

3.5　拒絶理由通知書

本節では，拒絶理由通知書について具体的に説明する．

拒絶理由通知書は，拒絶査定前に特許審査官が示す通知書のことであり，出願に対する類似引用文献を示して，拒絶理由を示す場合がある．

拒絶理由が示された記載中に，出願特許文書と引用特許文書の対比箇所が明示されている．この対比を利用し，語の類似データを収集することができると考えた．

拒絶査定通知書の記載例を図 3.7 に示す．

記載例

拒絶理由通知書

特許出願の番号　　特願○○○○－○○○○○○○
起案日　　平成○○年　○月　○日
特許庁審査官　　○○　○○　　○○○○　○○
特許出願人代理人　　○○　○○
適用条文　　第２９条第１項

この出願は、次の理由によって拒絶をすべきものです。これについて意見がありましたら、この通知書の発送の日から６０日以内に意見書を提出してください。

理由

（新規性）この出願の下記の請求項に係る発明は、その出願前に日本国内又は外国において、頒布された下記の刊行物に記載された発明又は電気通信回線を通じて公衆に利用可能となった発明であるから、特許法第２９条第１項第３号に該当し、特許を受けることができない。

記　　（引用文献等については引用文献等一覧参照）

・請求項　　1
・引用文献等　1
・備考
　・・・・・・・・・・・・・・・・・

＜引用文献等一覧＞
１．特開昭○○－○○○○○○号公報

＜先行技術文献調査結果の記録＞

・調査した分野　　ＩＰＣ　　Ｂ43Ｋ 8/00　～　8/24
　　ＤＢ名
・先行技術文献　　特開平○○－○○○○○○号公報
　　（本願の発明の詳細な説明中，明細書，段落○○○○，第○行に記載されている「Ｂ」の点については，本文献第○頁，第○欄，第○行に記載されている。）
・出願人への要請
引用文献１は、本願出願時に公開されており、本願と出願人又は発明者が共通する文献であって、本願の一以上の請求項について、当該引用文献のみで新規性又は進歩性を否定するものです。
このような文献に基づいて、事前に発明を適切に評価することは、出願人による適切な請求項の作成に役立つとともに、迅速かつ的確な審査にも資するものと考えられます。出願・審査請求の際には、このような文献を出願人が知っている先行技術文献として明細書中に開示するとともに、特許を受けようとする発明が、このような文献に基づき特許性を有するものであるか否かについて適切な評価を行っていただくようお願いします。

この先行技術文献調査結果の記録は、拒絶理由を構成するものではありません。

この拒絶理由通知の内容に関するお問い合わせ又は面接のご希望がありましたら次の連絡先までご連絡ください。
審査第【漢数字】部【審査室】　【審査官（補）名】
ＴＥＬ．０３－３５８１－１１０１　内線
ＦＡＸ．０３－　　－

図 3.7　拒絶理由通知書の記載例

図 3.7 の備考欄には，実際に，審査官が出願文献と引用文献を比較し，２つの対比箇所が書かれている．

図 3.8 に対比例を示す．

引用文献１（段落【００２９】－【００５２】、図１－３等参照）には、複数のサーバ（サービス提供装置）で発生するログデータを，前記サーバと同一企業が管理するログ管理装置（ログ収集サーバ）に収集し、該ログデータが外部の監査実行装置（ログ受信サーバ、データ蓄積サーバの機能を備えたサーバシステム）に送信され、分析されるログ監査システム（ログ収集システム）において、ログ管理装置（ログ収集サーバ）は、収集したログデータのログダイジェスト（第１ハッシュ値）を生成して、監査実行装置（サーバシステム）に該ログデータとログダイジェストの組を送信し、監査実行装置（ログ受信サーバ）は、受信したログデータのハッシュ値（第２ハッシュ値）を生成し、受信したログダイジェストと比較して、受信したログデータの改ざんの有無を検出することが記載されている。

図 3.8　特開 2009-053992 拒絶理由通知書より

図 3.8 のように特許審査官が，出願に対して審査官自ら調査した結果，類似の証拠として引用した文書の中に書かれている箇所と出願書類に書かれている箇所との対比を明確にして記載している．

対比を整理すると，下記図 3.9 のようになる．

サービス提供装置-----複数のサーバ
ログ収集サーバ-----ログ管理装置
ログ受信サーバ、
データ蓄積サーバの機能を備えたサーバシステム-----外部の監査実行装置
ログ収集システム-----ログ監査システム
第1ハッシュ値-----ログダイジェスト
サーバシステム-----監査実行装置
ログ受信サーバ-----監査実行装置
第2ハッシュ値-----ログデータのハッシュ値

図 3.9　拒絶理由通知書内の対比箇所

図 3.9 のように対比箇所を整理することができる.

そして，整理した対比箇所を類似語データとして収集可能であると考えられる.

3.6　拒絶理由通知書利用の効果

前節では，拒絶理由通知書の中には，特許審査官が最終判定した類似箇所が記載されており，これらのデータを類似文書検索に利用することの可能性を示した．

本節では，そのデータを利用するとどのような効果が得られるのかを，これまでの辞書を利用した検索方法との違いから考えを述べ，それを次章以降での実験検証における前提とする．

特許査定の段階での特許審査官の判定は絶対的なものであり，特許審査官が類似であると示した語については，通常の辞書を作成する者に比べて，信頼度や確実度の点において，はるかに優れており，いわば知識の結晶であるといえる．

検索精度の向上には高度な辞書整備が不可欠だが，実際には高度な辞書が簡単に作れるわけではない．1 つずつ人間が特許文書を読み込み，そのうえで，この語とあの語が類似しているなどと判定していく作業は非常に手間や時間のかかる作業であり，実際的ではない．

また，期限が関係なく作成できる通常の辞書作成と比べ，特許文書に出現する語は時代背景によっても大きく変わっていく．例えば，現代では，数十年前は新しい発明であったフロッピーディスクに類似する発明や，フロッピーディスクに代わる類似語について，調査することは少ない．技術の進化とともに，発明や，それにまつわる用語も進化あるいは退化していくため，類似特許検索では，常に時代に沿った新しい辞書が必要となってくる．

拒絶理由通知書は，特許権取得に対する特許法の要件を満たさない場合，特許審査官が，拒絶査定を行う前に出願者に向けて示す書類のことである（図 3.10）． そして，特許法 29 条第 1 項の新規性，あるいは 29 条第 2 項の進歩性の条件を満たしていないと判断する場合には，その根拠として，類似する文献を，引用文献として示さなければならないとしている．

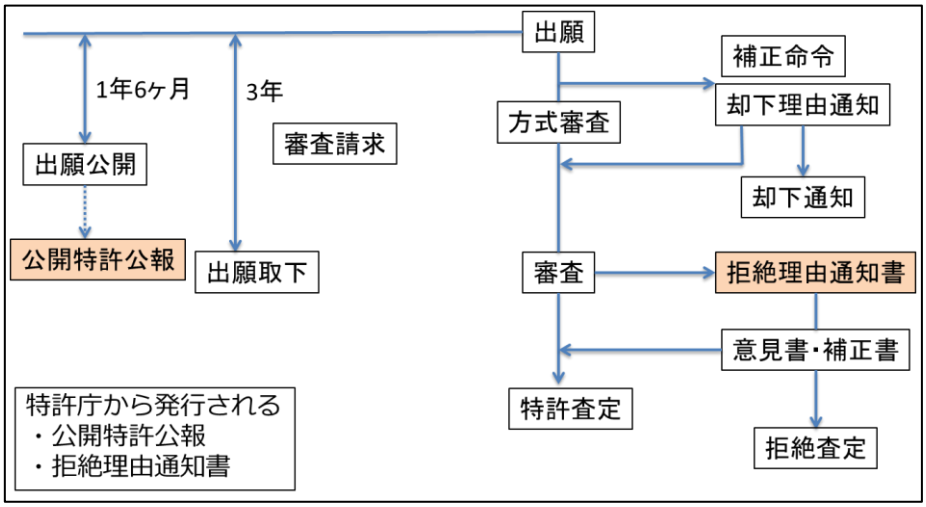


図 3.10　出願を起点とした特許審査のながれ

この拒絶理由通知書は，特許審査が行われている限り，半永久的に蓄積されていく．そのため，その時代にのみ使われていた語などの時代背景をくみ取ったデータの収集も可能となる．常に新しい語が自動的に蓄積されることになるため，人の目で確認をして，類似の判断を行わなければならないといった，もっとも判断の難しいところを特許審査官が担うといった形で信頼性の高いデータを得ることができる．

また，審査書類である拒絶理由通知書は，コストもかからず誰でも入手可能であり，審査の早期化も手伝い，知り得る情報を早い段階で入手することが可能である．

そのため，拒絶理由通知書から抽出した複合名詞を辞書として整備してまとめたデータを，類似複合名詞の辞書データとして利用することに高い価値があると考える．

3.7　結言

本章では，特許出願から査定までの流れと，特許を分類する国際特許分類番号について，そして，特許出願には欠かすことのできない先行技術調査の意義について説明した．

つづいて，出願者が先行技術調査を行う目的や方法について述べ，最後に本研究で利用する拒絶理由通知書について説明をした．

この拒絶理由通知書では，その記載内容から，特許審査官が最終判断した結果に基づき，類似とした語句を対比している．そして，整理した対比箇所は，類似語データとして収集可能であることから，複合名詞を利用した検索に有効に作用できるのではないかと考えられる．そして，拒絶理由通知書を利用することが特許文書類似検索において有益であるかについて説明すると共に，具体的な利用方法については，次章以降で述べていく．

第４章

拒絶理由通知書からの複合名詞抽出とその効果

特許文書検索において，重要なことは効率的な類似文書を見つけることである．効率的とは，再現性もあり，かつ適合性も高く，ノイズの少ない検索のことである．再現率と適合率は，トレードオフの関係により双方平均的な値で推移することになるが，ノイズの多い結果は非効率である．しかしながら，従来の形態素解析をベースとした類似文書を検索する方法を，特許文書に適用する場合，非常にノイズの多い結果となる．そのため，いかにノイズを拾わず，効率的に検索を行うかについて考える必要がある．

本章では，実際に形態素解析をベースとした検索は，非効率であるのかを検証する．そして，本論で主張したい複合名詞を利用した複合名詞解析の手法との比較を行う．つづいて，2 つの方法の検索結果より，形態素解析と複合名詞解析の問題点について説明する．次に，複合名詞解析について，どのようにして複合名詞解析の問題点について解決すべきかを述べ，最後に，複合名詞抽出について述べ，抽出した複合名詞を利用した複合名詞解析が特許文書検索に有効であることを示す．

4.1　形態素解析と複合名詞解析

一般的な文書で使われる形態素解析による文書類似度を測る方法と，複合名詞を利用した方法（以下，複合名詞解析という）を特許文書（ここでは請求項をさす）に適用する．

その結果，複合名詞の利用に際して，必要となる処理を確認する．

文書 A および B に含まれる文書中の要素，名詞を MP_{xi}，MP_{yi} とし，複合名詞を CN_{xi}, CN_{yi} としたとき，それぞれの要素をベクトル化し，コサイン類似度(cos_{mp},cos_{cn})を求める式を(式 4.1)および(式 4.2)に示す．

形態素解析　　$$cos_{mp} = \frac{MPxi \cdot MPyi}{|\overrightarrow{MPxı}| \cdot |\overrightarrow{MPyı}|}$$　　（式 4.1）

複合名詞解析　　$$cos_{cn} = \frac{CNxi \cdot CNyi}{|\overrightarrow{CNxı}| \cdot |\overrightarrow{CNyı}|}$$　　（式 4.2）

実際の類似度判断では，類似度の閾値を決めたうえで，類似の判断をすることになるが，ここではまず形態素解析をベースとした類似判断の問題点を明らかにするため，まずは機械的に判断する必要があった．そのため 1 つでも共通形態素があれば，類似と判定している．

4.1.1　対象文書の抽出

IPC 分類のうち，H，A，F，C の各セクションから 300 本合計 1,200 本の公開特許公報を抽出する．その中より，請求項 1 を抽出した．

請求項 1 を抽出した理由は，1 以下につづく請求項は請求項 1 に書かれるメ

インクレームの従属的な従属請求項であることがほとんどであり，発明の肝となる部分が請求項 1 に書かれているからである．

4.1.2　文書の下処理

各請求項 1（以下，文書という）の中で，請求項では，不要語とされている語を削除する．

＜削除した語＞
前記
当該
該
特徴
上記

4.1.3　解析の方法

抽出した文書を 2 つの手法で解析する．

1.　形態素解析

各文書を形態素解析する．形態素解析器は MeCab(形態素解析辞書 IPA-dic Ver.2.7.0)を用い，名詞を抽出した．

2.　複合名詞解析

各文書を複合名詞解析する．複合名詞抽出には，Term Extract を用いた．

3.　文書の類似度測定

・文書の類似度をコサイン類似度により算出する．その際に，複合名詞と形態素それぞれ抽出した語を単語ベクトルとした 2 種類の類似度を算出する．
・文書 No.1～No.1,200 までの文書を 1 つの文書に対して残りの 1,119 本の文

書と類似しているかどうかを確認する.

・正解文書の定義を同じセクションのものとし，類似の定義をコサイン類似度が 0 でないものとして，類似のありなしを判定した.

4.1.4　解析結果

セクションごとで，対象文書と類似ありと判定した文書が同じセクションであるかどうかで適合率，再現率を判断する．

これを，N0.1 から No.1,200 まで確認し，セクションごとの平均の適合率，再現率を示す．

表 4.1 は，文書 No.1 の例について，類似文書の適合率，再現率を示す．

表 4.1 文書 No.1 の適合率・再現率

文書No.1（セクションH）					
	類似文書数（全類似数1119）	セクション中の類似文書数（全類似数299）	再現率	適合率	F値
形態素解析	1184	299	1	0.25	0.4032
複合名詞解析	214	145	0.485	0.6776	0.5653

また，図 4.1 から図 4.4 は，セクションごとの形態素解析，複合名詞解析の平均適合率，再現率を示す．

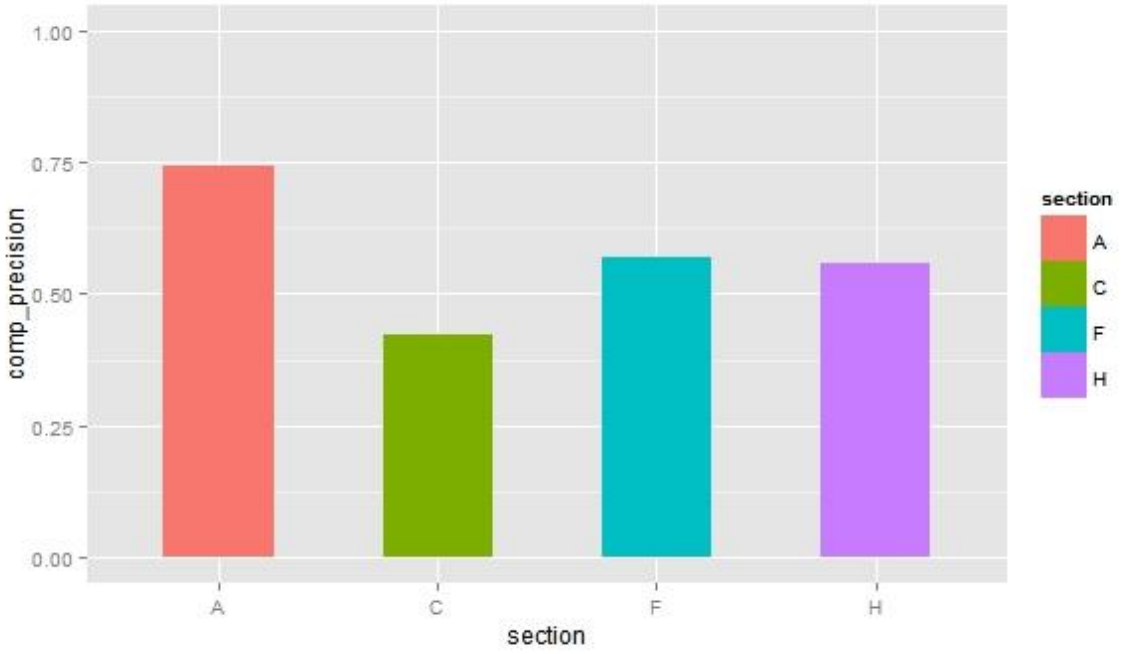


図 4.1 複合名詞解析の平均適合率

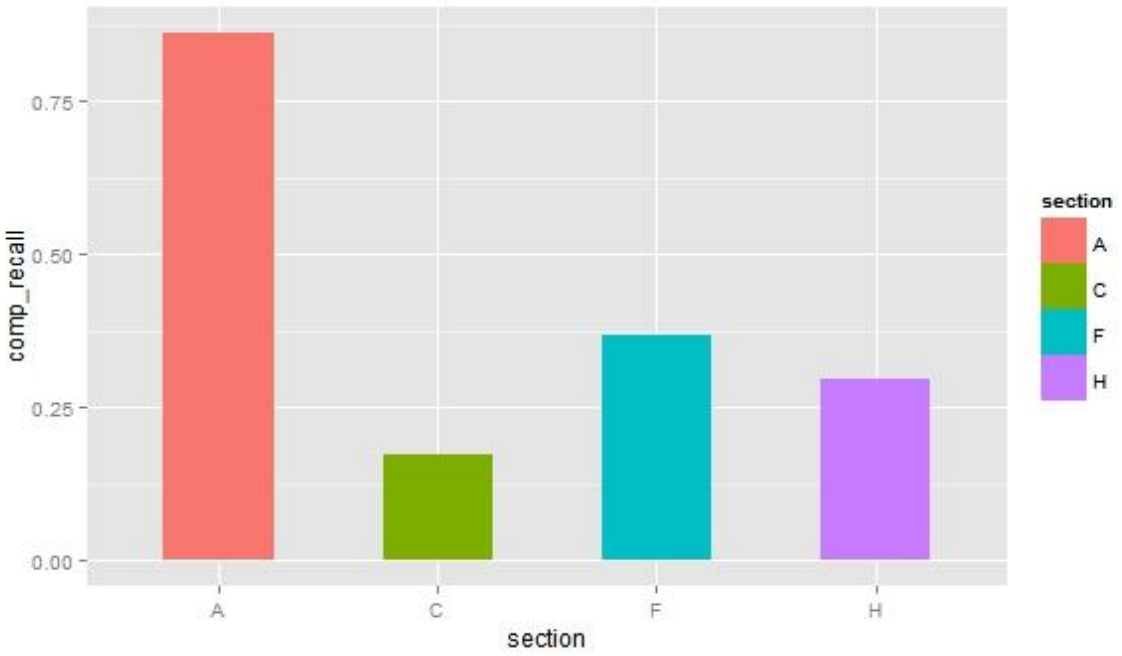


図 4.2 複合名詞解析の平均再現率

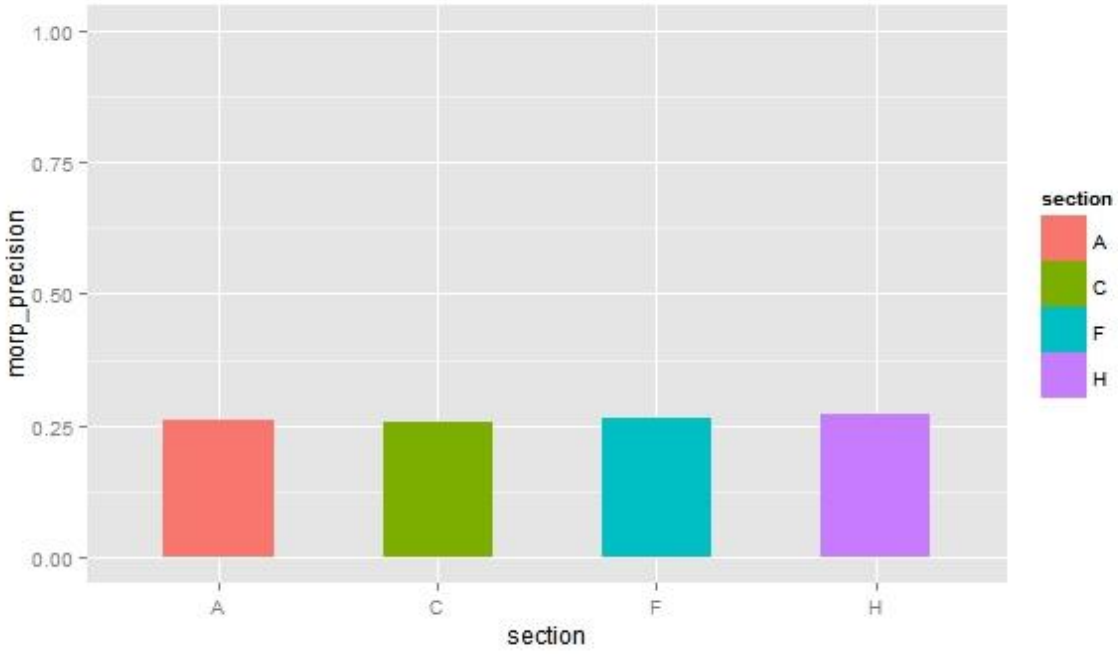


図 4.3 形態素解析の平均適合率

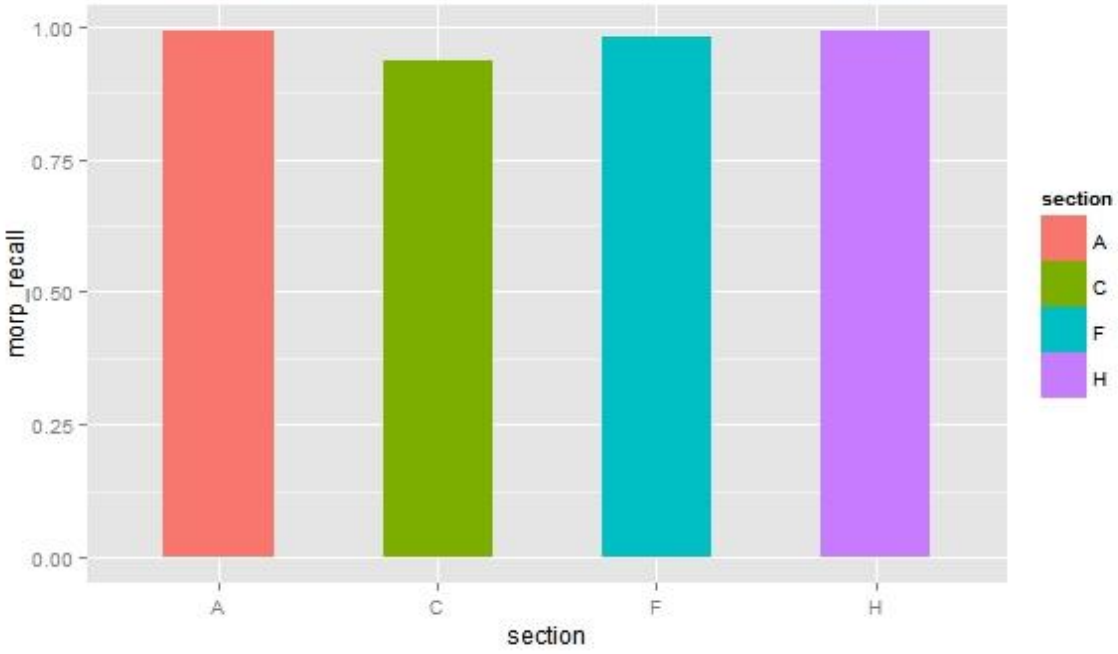


図 4.4 形態素解析の平均再現率

表 4.2 に結果をまとめる

表 4.2 各セクションの適合率・再現率

セクション	形態素解析 再現率	複合名詞 再現率	形態素解析 適合率	複合名詞解析 適合率	形態素解析 ノイズ率 %	複合名詞解析 ノイズ率 %
H	0.9925	0.2967	0.2697	0.5600	72.4932	33.4967
A	0.9925	0.8623	0.2612	0.7428	73.6230	22.8779
C	0.9359	0.1732	0.2547	0.4231	74.0891	44.0067
F	0.9818	0.3661	0.2637	0.5704	73.4974	35.0428

表 4.2 より，形態素解析は，再現率が高いが適合率が低い．この結果，本来は類似ではないノイズ（無駄な文書）を類似と判定してしまっている結果であり，類似文書検索において，形態素は類似の範囲が広くなりすぎるので精密な検索に向いているとは言えず, 非効率的であると言わざるを得ない（図 4.5 左）.

一方の複合名詞解析では，よい適合率を得られている．ただし，再現率が下ってしまっているので，この再現率をあげる工夫が必要となる（図 4.5 右）.

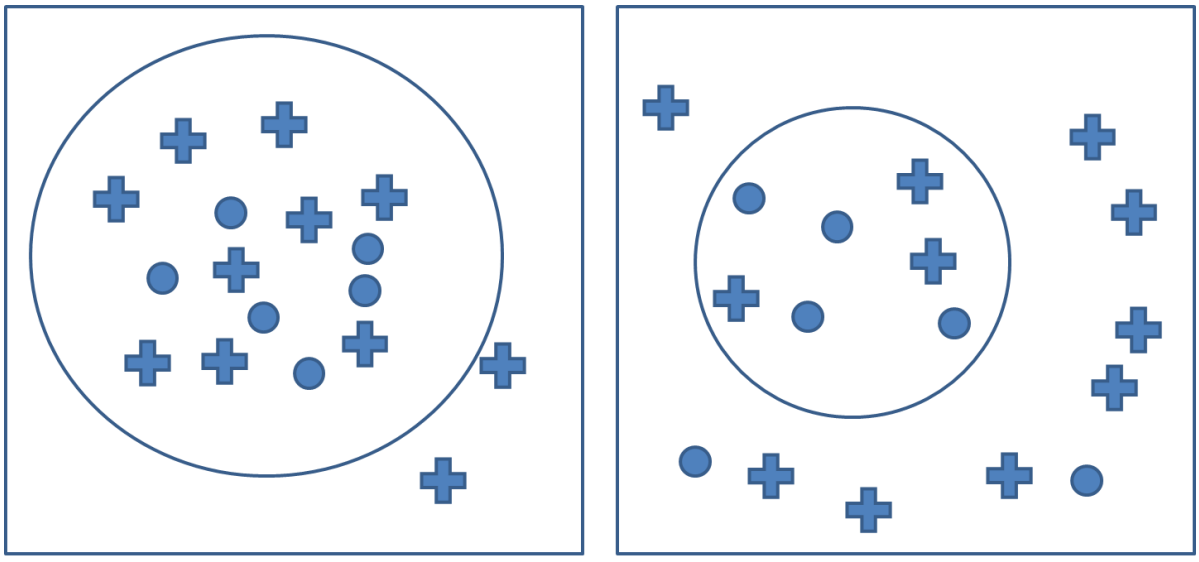
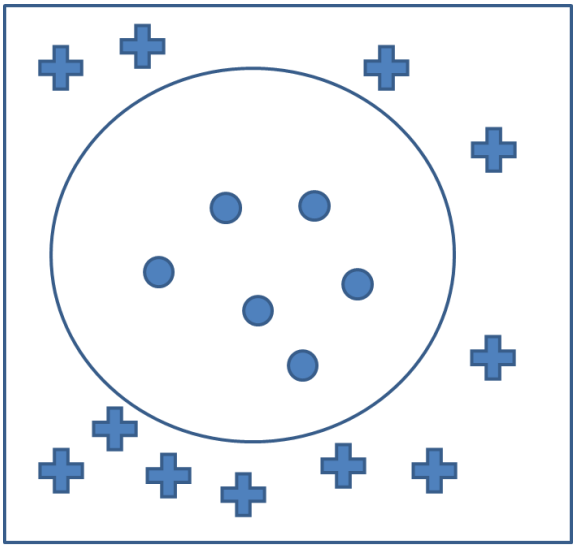


図 4.5　アンバランスな再現率・適合率

（● 正解文書 ✚ 不正解文書）

ノイズのほうが, 正解文書よりも多い場合, ノイズを除去することが難しい. 検索の理想形は，図 4.6 のような状態になることである.



類似文書検索の理想形

図 4.6　再現率・適合率の理想形

複合名詞解析の，再現率をあげるためには，同じ複合名詞の出現数を増やさねばならない．1 つの文書内に出現する複合名詞が，別の文書に出現する割合は，名詞だけの出現率に比べて低くなり，同じ複合名詞は出現しにくいかもしれない．しかし，類似する複合名詞がわかれば，それを利用し，複合名詞の類義語をキーワードとして文書の再現性をあげることができると考えられる.

4.2　複合名詞解析の問題点

本節では，前節で示した形態素解析と複合名詞解析を利用した文書分類の結果，複合名詞解析の問題点を明らかにする．その理由を示し，問題点についてどのようなアプローチが必要になるかを次節で述べる．

審査官が類似であると示した下記の 2 つの公報，審査対象の出願（以下，本願という）2010-094649 と，審査対象の類似文献として審査官が認めた文献（以下，引用という）2007-90218 を比較した．

本願

【公開特許公報】2010-094649

【発明の名称】バブリング装置

【IPC 番号】

B01F　3/12…混合，例．溶解，乳化，分散

B01D 21/02…分離

C02F 1/00…水，廃水，下水または汚泥の処理

【請求項 1】

不溶性成分を含む排水が貯留される排水貯留槽と，前記排水貯留槽に貯留される排水が流入し，排水に含まれる前記不溶性成分を沈殿させて除去する沈殿槽と，を有する，排水処理設備に備えられ，気泡を供給することにより前記排水貯留槽に貯留される排水中の不溶性成分の沈殿を抑制するバブリング装置であって，

前記バブリング装置は，前記排水貯留槽の底部に配置されるバブリング配管と，前記バブリング配管に空気を供給する**空気供給装置**と，を有し，前記バブリング配管は，前記空気供給装置からの空気を流通させるための主管と，前記主管に垂直に接続されて鉛直方向に伸び，それぞれに少なくとも一の空気噴出孔が

形成された，複数のノズルと，を有することを特徴とするバブリング装置．

引用

【公開特許公報】2007-90218

【発明の名称】有機性排水処理方法ならびに有機性排水処理設備

【IPC 番号】

C02F　3/12…水，廃水，下水または汚泥の処理

C02F　3/22…水，廃水，下水または汚泥の処理

C02F　1/44…水，廃水，下水または汚泥の処理

B01F　7/16…混合，例．溶解，乳化，分散

B01F　3/04…混合，例．溶解，乳化，分散

B01F　5/10…混合，例．溶解，乳化，分散

【請求項 1】

活性汚泥により有機性排水を曝気槽で好気的生物処理する有機性排水処理方法であって，

前記曝気槽あるいは前記曝気槽よりも下流側の活性汚泥を，酸素を含有する気体とともに**超微細気泡発生装置**に導入して前記活性汚泥中に含有されている活性汚泥凝集体を微細化させるとともに活性汚泥中に前記気体を１００μｍ以下の微細気泡として形成させ，該微細気泡を含有する活性汚泥を前記曝気槽に返送して前記微細気泡の散気を実施することを特徴とする有機性排水処理方法．

双方ともジェプソン形式で書かれており，発明の新規性にあたる部分は下線を引いてある．この部分を中心に 2 つの文書を比較した．

まず，キーワードで検索してみる．検索には，独立行政法人 工業所有権情報・研修館の特許データベースを用いた検索システムを利用した．

本願のタイトルである「バブリング装置」では，引用を検索することができ

なかった．次に，本願のもつ IPC 番号で分類検索をしたが，引用を検索することができなかった．そして，本願下線部で出現した複合名詞「空気供給装置」をキーワードとして検索したが，やはり引用を検索することができなかった．

しかし，請求項から複合名詞抽出した結果より，複合名詞の中で「装置」という同じ語が使われている複合名詞「超微細気泡発生装置」に着目し，これを「空気供給装置」と類義として，類義語で検索を行ったところ，正解文書にあたることができる（表 4.3）．

表 4.3　「空気供給装置」の検索

検索方法	検索語のみ	IPC 検索式	類義語を利用
	「空気供給装置」	(B01/F3/12)or (B01/D21/02)or (C02F1/60)	「超微細気泡発生装置」
検索結果	424 件	1031 件	13 件
検索結果中の 正解文書の有無	×	×	○

つまり，本願と引用を類似文書と判定するために，検索に用いる 2 つの複合名詞「空気供給装置」と「超微細気泡発生装置」が類似する複合名詞であるとするデータベースが必要になると考えられる．

4.3　複合名詞抽出の方法

前節で，類似する複合名詞のデータベースが必要であることが分かった．一方，出願審査の過程で特許審査官示す「拒絶理由通知書」に着目したことから，本願に対して引用をもち，審査官が類似理由を対比して明示している箇所を抽出し，本願と引用の類似複合名詞データを表 4.4 にまとめた．

IPC 番号 H04M11/00 をもつ公報で，拒絶理由に特許法 29 条第 1 項 3 号の新規性の要件を満たさずに拒絶査定となったものを抽出対象とした．

表 4.4　類似複合名詞データ

本願	引用
外部デバイス	外部端末
撮影報知手段	撮影予告信号
監視データ発生時刻	異常発生時間情報
携帯電話発信抑制装置	遠隔制御装置
コンテンツデータ	広告情報
利用開始情報	サービス提供予想時刻
文字列画像	文字行記載領域
ネットワークノード	サービス提供機器
音声ファイル化リンガー	メロディデータ
転送接続	着信転送制御手段
電話帳管理手段	個人情報管理方式

優先呼制御手段	FAX 通信時機能制限手段
リサイズ手段	再符号化
サーバ側記録部	個人管理用データベース
移動体通信端末装置	携帯用電話機
情報公開システム	情報提供システム
位置情報	GPS 情報
携帯情報端末装置	移動体
携帯端末装置	携帯情報端末
選択制御手段	通信方式選択手段
体調データ	バイタルデータ
監視カメラ動作機能	CCDカメラ制御ユニット
発呼先	着信側携帯電話
画像データ蓄積手段	画像保存手段
文字行記載領域	文字画像
マルチメディアメッセージ	電子メール
防災受信機	アナログ感知器
通信網側	インターネット
名称情報	ニックネーム
ゲートウェイ装置	シグナリング中継システム
情報提供システム	コンテンツ配信方法

コンテンツ	音楽データ
ホームネットワーク機器	制御対象機器
解決支援情報	制御情報
イベント検出手段	エラーコード取得手段
設備機器監視装置	サポート装置
携帯端末	顧客端末
分散型電源設備	プラント
分散型電源設備	プラント
監視情報	プラント運転データ
ハンドヘルド装置	子機
着信情報	取り次ぎ管理情報
対象物指定部	位置特定情報
対象物指定部	空間範囲情報
許可パケット	送信許可情報
画像変換	データ変換後文書
受信能力情報	能力有無判断手段
通信路判定部	ネットワーク混雑度測定手段
制御信号	運転情報
一般公衆網	公衆電話網インタフェース
情報送信システム	データ管理システム

配車要求データ	車両要求情報
2次元バーコード	符号化画像
移動通信端末	携帯型通信機器
移動通信端末	移動体通信装置
メール保存システム	メール格納処理手段
被監視局	被制御所
各種認証用データ	相互接続装置識別番号
各種認証用データ	本人属性情報
撮影可能全範囲	可動映像
詳細広告送出手段	記事返送手段
送出手段	記事返送手段
広告蓄積手段	広告管理手段
画像データ	撮像画像
サービス品質情報	通信品質レベル
サービス品質情報	要求 QoS
住人異常報知手段	人体異常信号
音声再生データ	デジタル音声信号
温度検出手段	温度検出回路
遺失物回収方法	紛失物返還支援方法
集約ユーザ嗜好データ	ユーザ嗜好情報

ソフトタグ情報	情報パッケージデータ
制御信号	機器命令列
送信元送信手段	メール配信システム
所定時間開錠状態	施錠/開放機構
識別情報付加手段	宛先別メール配信手段
情報通信端末	情報交換端末装置
監視対象	監視地点
異常発生	特定事象
画像データ	オブジェクトデータ

請求項の文字数は日本語だと 300 文字程度のものが多いので，少ない情報量の中から効率的な検索結果を得るために，表 4.4 のデータを類似複合名詞の辞書データとして利用することができればよいと考えられる．

ただし，抽出したデータすべての複合名詞を類似とすることはできない．なぜなら，抽出した複合名詞の中には，「第一部材」，「所定位置」など，抽出元文書にだけ特有な複合名詞も存在するからである．

そのため，次には抽出した複合名詞をどのように同定するかが課題となる．どの複合名詞とどの複合名詞が類似あるいは同義または概念一致（上位概念・下位概念）であるかを考慮した類似の複合名詞辞書構築をしなければならない．

4.4 結言

本章では,類似文書検索の一つの特徴である複合名詞解析について提案した.そして，特許拒絶理由通知書から抽出することのできる類似複合名詞の辞書データ作成の可能性を示した.

はじめに，一般的な類似文書を検索するときに用いられる形態素解析と比較することで，ノイズが多く適合率が低くなってしまう形態素解析手法よりも複合名詞解析手法を利用した方が，効率的な検索が行えることを示した.

次に，複合名詞を利用した解析を行う場合の複合名詞解析の問題点を明らかにした．問題は，複合名詞は，形態素より特許文書の特徴を捉えられるが，文書に出現する頻度が少ないため，通常利用される文書類似計算の手法（コサイン距離など）での類似の数値は低くなってしまうという点である.

また，複合名詞を類似特許検索の際のキーワードに定めることの問題点は，人の目で見て類似であろうと判断できるものであっても，機械が非類似と判断してしまうことである．例えば，複合名詞を構成する名詞の語順の違いなどでも機械的に非類似と判断されてしまう場合がある．このように比較する文書中に同じ複合名詞が出現する確率は非常に少なくなり，結果として類似する文書を正しく探しだすことができない.

この問題点を解決するために，類似となる複合名詞の意味拡張することを考えなければならない．複合名詞は出現する確率が形態素に比べて少ないため，ある程度の類似の幅を拡張しなければならず，そのための辞書の構築が必要である.

本研究では，類似する複合名詞を探すために，特許審査において審査官が拒絶査定する前に示す，拒絶理由通知書を利用できるということを発見した.

この拒絶理由通知書の中で，特許法 29 条第 1 項を満たしていないことを理由とした場合は，審査官が引用文献を示すため，出願書類と引用文献で対比された複合名詞を双方から抽出することができる.

ただし，文書固有の複合名詞などもあり，抽出したすべてが類似と判定され

るものではない場合もある．この対比群のうち高い精度で類似を示すことができれば，分野ごとに類似辞書として活用できることが考えられる．

キーワード検索をする場合であっても，1 つの複合名詞だけでなく 2 つの複合名詞を使って検索することで，検索範囲の幅を広げることができるため，辞書を作成することは有効であるといえる．

それゆえ, 作成した辞書を利用することで, 文書検索を行った際にそれまで, 探し出せなかった文書を類似文書として探し出せる可能性が高まる可能性も示唆できる．

ただし, それまで拒絶理由通知書から複合名詞を抽出した活用例がないため, 抽出した複合名詞が正しく利用できるかどうかを検討する必要がある. 次章で, その検討を行う．

第５章

複合名詞の類似度判定手法

前章で行った複合名詞の抽出手法について，抽出した複合名詞を辞書データとして機能させるため，複合名詞の類似度判定の検証を行った．なぜならば，抽出した複合名詞がすべて類似として判断されるわけではないからである．

本章では，複合名詞の類似度判定手法について述べる．まず，構文解析手法から判定する方法について述べ，次に，複合名詞を構成する各形態素の類似から判定する方法について示す．さらに，複合名詞の類似分類評価の内容と結果について述べ，最後に，複合名詞の類似評価を行い，その結果，類似判定手法が複合名詞の類義語辞書構築において必要であることを示す．

5.1　複合名詞を類似判定する方法

前章で，拒絶理由通知書から抽出した複合名詞が，類似語辞書として類似特許文書検索に利用できる可能性を示した．ただし，抽出した語がすべて辞書として利用できるものではない．そのため，本節では，抽出した複合名詞の類似判定をする手法を示す．

特許文書のなかの定型表現（としての，からなる，を備えた，を用いた，を記述した）から語句関係を整理し，上位概念と下位概念の類似をみつける方法がある[Uchiyama 07]．例として，「単語列 からなる 出力文」「意味情報 を記述した 単語辞書」という文より，単語列と出力文，意味情報と単語辞書は，それぞれ上位概念・下位概念としての類似複合名詞と判定する．ただし，このような方法だと，定型表現を介していない複合名詞は類似判定ができない．本節では，従来方法のような，定型表現にはとらわれず，複合名詞を類似判定する方法について述べる．

5.1.1　構文解析手法の利用から判定する方法

抽出した複合名詞の類似を判定する．まず，文の構造から考える構文解析手法を考える．

構文解析（係り受け解析）は，特許文書のような長い係り受け文が何重にも係っていると，エラーとなりやすい．一般的な係り受け解析器は新聞などの整った文の解析に強いが，ウェブ上にあるブログなどの整っていない文の解析に関しては難しいとされている．そこで，ウェブ上にあるブログなどの文での解析精度が高いとされる Yahoo API の係り受け解析を利用した（http://developer.yahoo.co.jp/webapi/jlp/da/v1/parse.html）．

まず，図 5.1 で類似語判定のながれを確認する．

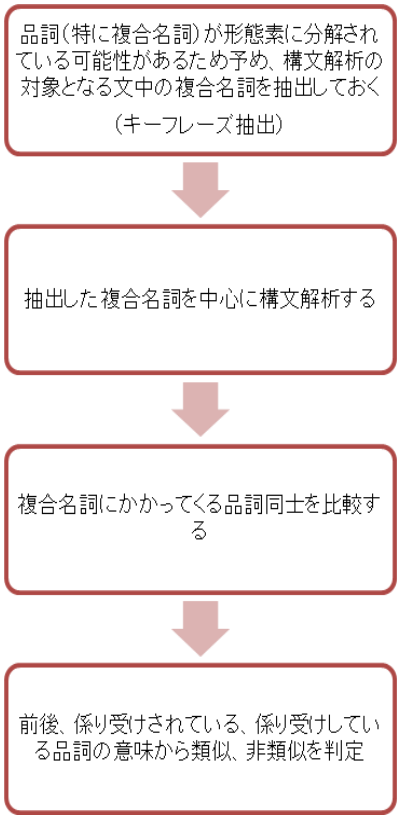
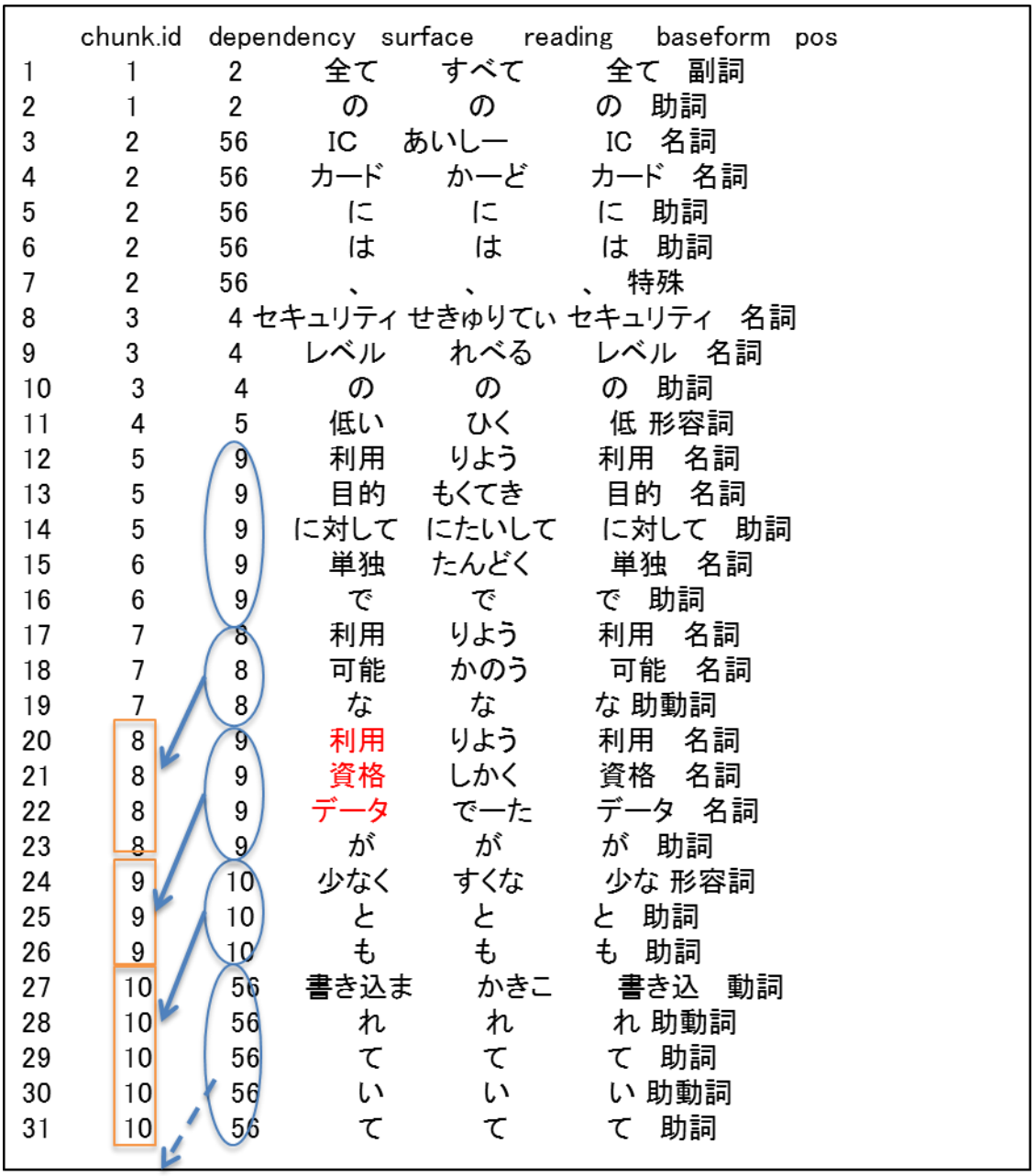


図 5.1　係り受け解析を利用した類似語判定

注目すべき，複合名詞を含む文節に注目し，係り受けからの類似を判断する．本願と引用において，対比するそれぞれの複合名詞に係る部分と係り先を比較する．



	chunk.id	dependency	surface	reading	baseform	pos
1	1	2	全て	すべて	全て	副詞
2	1	2	の	の	の	助詞
3	2	56	IC	あいしー	IC	名詞
4	2	56	カード	かーど	カード	名詞
5	2	56	に	に	に	助詞
6	2	56	は	は	は	助詞
7	2	56	、	、	、	特殊
8	3	4	セキュリティ	せきゅりてぃ	セキュリティ	名詞
9	3	4	レベル	れべる	レベル	名詞
10	3	4	の	の	の	助詞
11	4	5	低い	ひく	低	形容詞
12	5	9	利用	りよう	利用	名詞
13	5	9	目的	もくてき	目的	名詞
14	5	9	に対して	にたいして	に対して	助詞
15	6	9	単独	たんどく	単独	名詞
16	6	9	で	で	で	助詞
17	7	8	利用	りよう	利用	名詞
18	7	8	可能	かのう	可能	名詞
19	7	8	な	な	な	助動詞
20	8	9	利用	りよう	利用	名詞
21	8	9	資格	しかく	資格	名詞
22	8	9	データ	でーた	データ	名詞
23	8	9	が	が	が	助詞
24	9	10	少なく	すくな	少な	形容詞
25	9	10	と	と	と	助詞
26	9	10	も	も	も	助詞
27	10	56	書き込ま	かきこ	書き込	動詞
28	10	56	れ	れ	れ	助動詞
29	10	56	て	て	て	助詞
30	10	56	い	い	い	助動詞
31	10	56	て	て	て	助詞

図 5.2 係り受け解析の結果

図 5.2 では，請求項を係り受け解析した．キーフレーズとなる複合名詞を抽出しているので，複合名詞が形態素に分解されてしまっていても，1 つのかたまりとして捉えることができる．

キーフレーズとなる複合名詞「利用資格データ」を含む文節の係り受けを図

5.3 で確認する．

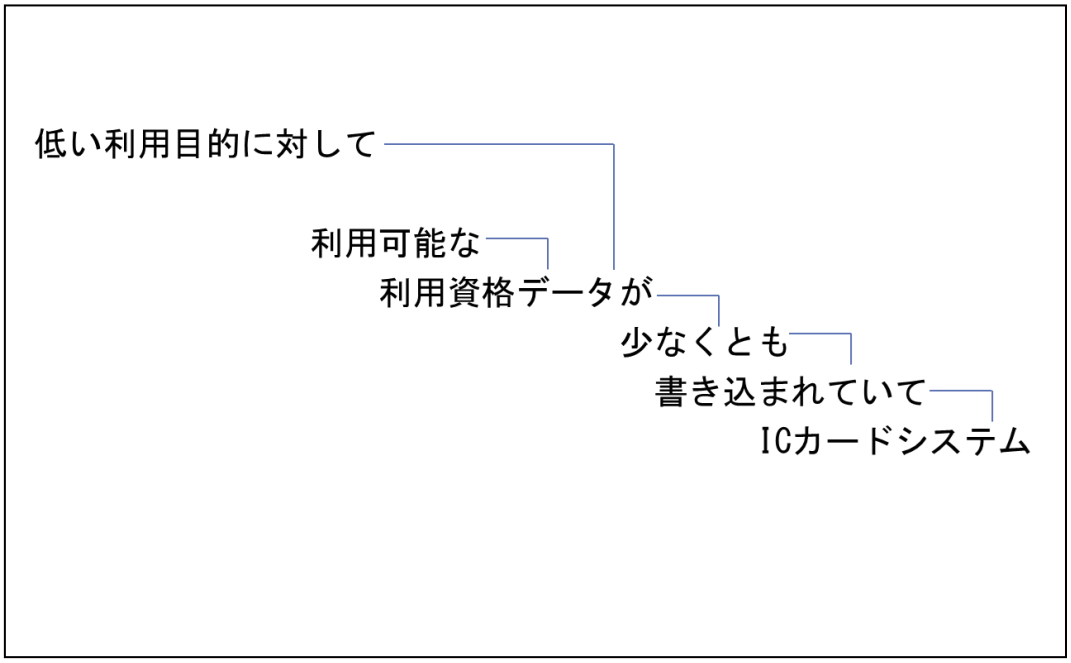


図 5.3　文の係り受け

この例で，「利用資格データ」は「書き込まれていて」つまり「書き込む」という動詞に係っている．また，「書き込む」の係り先には別の複合名詞「IC カードシステム」が出現している構成となっている．

そして，この請求項の引用となった一文を図 5.4 で比較する．

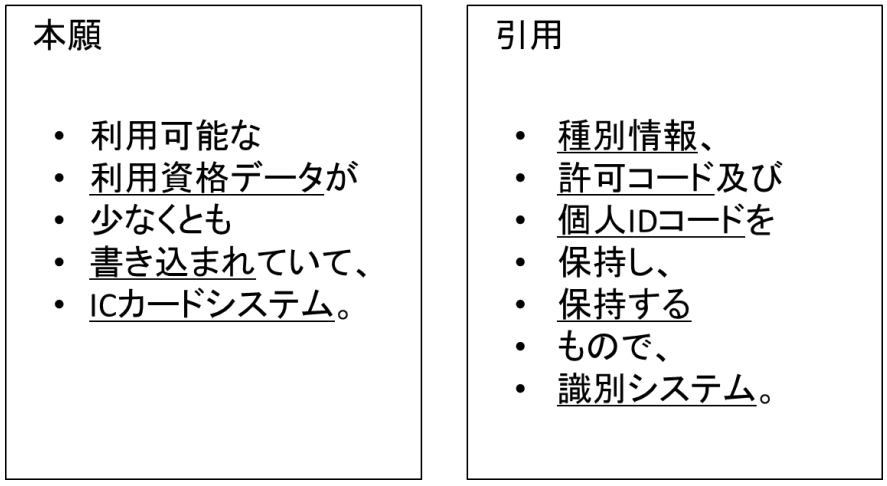


図 5.4　本願と引用の文の比較

類義語辞書により，類似であるとわかっている「書き込む」 と 「保持する」の 2 つの動詞の前後に出現する複合名詞を比較することができる．

これによると，「利用資格データ」と「種別情報」「許可コード」「個人 ID カード」が類似，そして，「IC カードシステム」と「識別システム」が類似と判定できる．

5.1.2　複合名詞内形態素の類似から判定する方法

複合名詞を構成する名詞を形態素解析する例を下記に示す．分解された形態素（名詞）を比較することで，構成された複合名詞が，文字列照合，同義語・類義語，シソーラスどの分類に当てはまり，類似となったのかを判断する．

・文字列照合
温度/管理/室≒温度/制御/室

文字列照合は,形態素の文字が同じであることをいう(温度＝温度,室＝室)．

・同義語・類義語
温度/管理/室≒気温/コントロール/ルーム

同義語・類義語は，気温と温度が同じ意味で使われる，あるいは，管理の英訳がコントロールとなることなどで，まったく同じ文字列ではないが，この場合は同義語あるいは類義語にあたるとする．

・シソーラス
温度/管理/室≒冷蔵/庫

シソーラスについては表 5.1 の上位概念，下位概念の考え方を用いる．
シソーラスとは，単語の上位と下位関係，部分と全体関係，同義関係，類義関係などによって単語を分類し，体系づけたものである．
表 5.1 では，シソーラスの例を示す．

表 5.1　シソーラスの例

	意味的分類	意味的分類	言葉的分類
上位概念	柑橘類	果物	車
下位概念	みかん	バナナ	牽引車

図 5.5 は，実際の拒絶理由通知書であり，括弧内の複合名詞は，本願で使われている語で括弧の直前に書かれている複合名詞が引用で使われている語である．

引用文献1に記載された発明においては、ICカードリーダライタ(読取装置)は、セキュリティレベルの低い利用目的(園児の入場)の場合は、単独のICカード(IDカード)に書き込まれた利用資格データ(許可コード)が読み取れたことで利用を許可し、セキュリティレベルの高い利用目的(園児の退場)の場合は、複数枚(2枚)から成る1組のICカード(IDカード)のうちの1枚に書き込まれたID番号と、これと関連付けられた他のICカードのID番
号とがそれぞれ読み取られて、互いに関連した複数枚(2枚)から成る1組のICカード(IDカード)であることが判明したときに利用を許可するように構成されている。

図 5.5　拒絶理由通知書一文

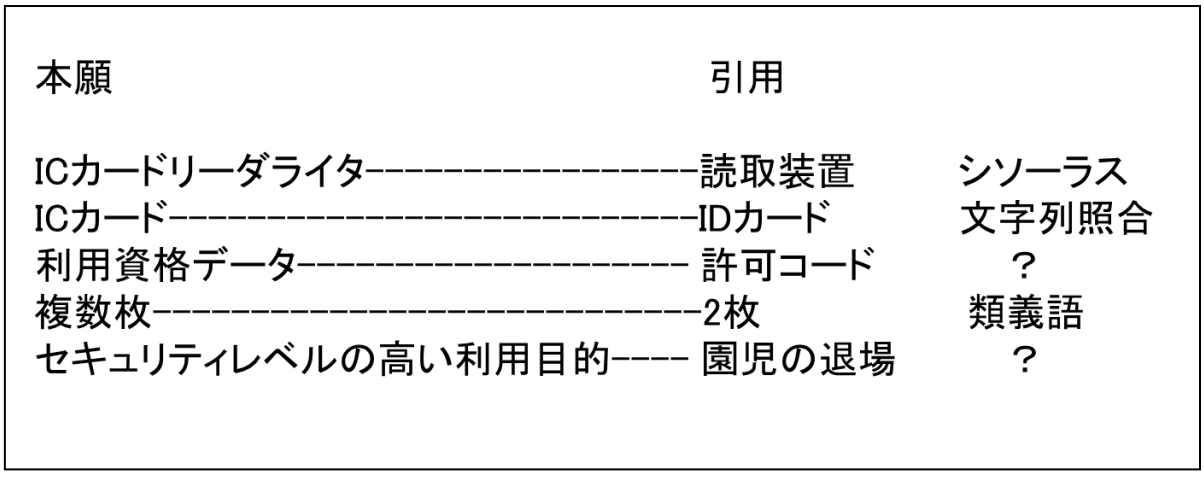


図 5.6　本願・引用中の複合名詞の対比

図 5.6 は，図 5.5 の文から対比を抽出したものである．

ここで対比された複合名詞の比較が，文字列照合，同義語・類義語，シソーラスどの分類に当てはまるのか，それとも当てはまらず，非類似として辞書データとしては利用できないという判定とするのかを判断する．

本手法の利点は，複合名詞を構成する形態素（名詞）の語順が違っても類似を判定できることである．

個々の形態素の一致については，既存の類義語辞書，シソーラス辞書を利用した[Kitahara 01][Nakamura 05][Shibata 02][Yamaguchi 06][Oono 81]．

判断の具体的手法については，5.3 節で述べる．

5.2　複合名詞の類似評価

複合名詞を抽出し，5.1.1 および 5.1.2 の手法で複合名詞の類似を判定し，辞書データを作成した.

対象データ：
公開特許公報番号 2003-000001～2003-007500：　7500 件
29 条 1 項を理由に拒絶されている公報数：　137 件
対比明らかな複合語対：453 組

文字列照合では 179 件の合致で 39.51%，そこに同義語・類義語判定を加えると 221 件の合致となり 48.79%, シソーラス判定で 311 件の合致で 68.65%，最終的に，係り受けの解析を加え 366 件で 80.80%は類似となった.

抽出した対比の中には，複合名詞ではなく，単名詞もあるが，ここでは，形態素数 1 の複合名詞として扱う.

上記より，審査官が類似とした複合名詞のうち，8 割程度は類似の語として扱うことができるとわかった.

この複合名詞の類似評価では，とくに分類を指定せず，公開特許公報の番号順に抽出した.
結果を表 5.2 に示す.

表 5.2 複合名詞類似評価

No.	本願	引用	文字列照合	同義語類似語	シソーラス	係り受け解析

1	ミッションケース	ミッションケース M	○			
2	支持部材	後部連結フレーム８５				○
3	旋回モータ	モータ１５	○			
4	検知体	スイッチ3a			○	
5	脱穀部	扱胴１・扱室２等				○
6	排藁チェン	株元搬送体１８				
7	センサ部材	センサープレート 61b		○		
8	センサ部材	排藁処理感知部 62				
9	管理装置	ペット宿舎システム				○
10	識別手段	IC チップ			○	
11	管理サーバー	データベース			○	
12	携帯情報端末	飼い主用端末	○			
13	表示手段	液晶表示部４３	○			
14	情報獲得手段	データ線４９				
15	フランジ	上面開口周縁				
16	ガスケット	マグネット入りパッキン			○	
17	風筒	送風筒	○			
18	プラスチック基板	高分子ゲル				
19	損傷検出手段	光検出部	○			
20	検出結果解析手段	データ処理部				○
21	プローブ	DNA 鎖切断現象の検知部			○	
22	酵母由来の不溶性画分	酵母細胞壁画分	○			
23	シート	シート体	○			
24	バッグ本体の周壁部	バッグ本体	○			
25	底面	側面体				○
26	熱可塑性合成樹脂帯状片	接合体				○
27	第 1 部材	表面板 2				
28	第 2 部材	取付部材 4				
29	連結部材	ネジ 9			○	

30	指掛け部	指係止部 14	○			
31	治療器具	健康増進具			○	
32	係合部	ベルト連結片				
33	突起部	リブ				
34	挿入孔	ベルト挿入孔	○			
35	挿入溝	嵌合凹部			○	
36	粘着性材	粘着部材	○			
37	遠心力測定器（2）	スイング強度測定装置	○			
38	表示部（3）	強度測定メーター				
39	指示部材（6）	指針				
40	引き寄せスプリング	バネ			○	
41	位置保持機構	ラック				
42	リセット手段（7）	復帰ボタン			○	
43	前扉	開閉前枠 3			○	
44	意匠部材	透光性絵柄板 7a			○	
45	光源	照明灯装置 8a		○		
46	ブラケット	押え枠 11a		○		
47	特定領域部	V ポケット				
48	玉転動部	ジュータン				
49	ペットロボット	挟み上げ装置 13				
50	係合孔	貫通孔 11a	○			
51	係止ピン	クリップ部 5		○		
52	固定基台	支持台 23		○		
53	ケーシング	胴体 4				
54	排気管	出口側内管 13			○	
55	旋回羽根	旋回ら旋羽根 9	○			
56	入口	入口フランジ 1	○			
57	孔	開口部"G"			○	
58	出口	出口フランジ 2	○			

59	旋回室	三次ドレン分離室 8	○			
60	排液口	ドレンポット 16		○		
61	ガス吸収部	充填塔				○
62	ガス選択吸着材料	オルガノシロキサン重縮合物				
63	光触媒層	酸化チタン層			○	
64	疎水性の環境ホルモン	ノニルフェノール			○	
65	疎水性磁性粒子	吸着組成物				○
66	処理対象原水	排水			○	
67	酸化剤	圧縮空気			○	
68	流動領域	巡回流			○	
69	粒状酸化触媒の供給部	隔壁 3				
70	アルカリ剤	水酸化ナトリウム			○	
71	無機物	MAP 粒子			○	
72	線材供給装置	給線部				○
73	加工機	折曲加工部	○			
74	線材排出装置	排出部	○			
75	切断装置	切断部	○			
76	補修材	溶接材料				
77	切粉捕捉ケース	防塵カバー				
78	遮蔽部材	研磨粉受け 6				
79	加工中断指令手段	割込み信号			○	
80	中断処理手段	演算制御部				
81	基準値	スレショルド値	○			
82	外力追従機能	ダイレクトティーチングモード			○	
83	基材	基材 2	○			
84	着色したシート	着色紙 4	○			
85	突き板	極薄化粧単板 1	○			
86	クリアー塗装	透明塗料層	○			
87	衝撃制御装置	弾性棒体				

88	揺動テーブル	可動支持体				
89	揺動機構	回転駆動部材				
90	プラスチック添加剤	防蟻剤				○
91	ポリスチレン系発泡樹脂	スチレンビーズ			○	
92	非晶性プラスチック材料	ポリカーボネート			○	
93	微細加工面	凹凸形状				○
94	無機化合物	カーボンブラック			○	
95	板状物	第 1 のレンズ及び第 2 のレンズ				○
96	樹脂製部材	ランプケース 3				○
97	接着剤層	ハードコート層			○	
98	転写支持体	プラスチックフィルム				○
99	機能性微粒子	導電性粉末				?
100	機能性微粒子含有層	熱硬化型導電層				○
101	支持体	プラスチックプレート				○
102	ジョブ処理情報管理サーバ	POS サーバー	○			
103	着脱可能な記憶媒体	smart card			○	
104	凹溝	内周溝 6	○			
105	排出部材	排出フィン	○			
106	細リブ	偏摩耗防止用陸部 5				○
107	突起部	突起 4e	○			
108	始動点装置	始動点制御子	○			
109	終動点装置	終止点制御子			○	
110	踏切遮断照査器	踏切制御装置	○			
111	通信ネットワーク	伝送ネットワーク	○			
112	車輪部分	キャスタ 23,24		○		
113	運搬車	移動台車 1		○		
114	運搬車	手押し台車 1		○		
115	物品収納用ボックス	収納箱型本体 3	○			
116	物品収納用ボックス	コンテナ本体 2		○		

117	走行トランスミッション 基本伝動機構	前半動力伝達経路 a 及び左右後半動力伝達経路 b				
118	差動変速機構	操向変速装置 B	○			
119	緩旋回	緩旋回状態	○			
120	ゼロ旋回	信地旋回状態	○			
121	逆転旋回	超信地旋回状態	○			
122	ホッパ部	張込漏斗 14		○		
123	移送部	バケットコンベア 3			○	
124	シュート部	排穀樋 4				
125	排出管	送穀樋 8				
126	オーバーフロー管	穀粒溢流管 7			○	
127	シャッター	切替弁 11				
128	媒体	紙幣				○
129	収納部	金種別カセット				○
130	送出手段	分離モータ				○
131	取出手段	搬送モータ				○
132	媒体取出装置	紙幣入出金機				○
133	封入物	シート片 4A～4D				○
134	第 1 搬送部	支持板 1				○
135	第 2 搬送部	帯状材 11B，11C，11D				○
136	石英ガラス溶融炉本体	るつぼ（35）			○	
137	成形用ノズル	オリフィス（40）		○		
138	ダミー部材	おとり部分		○		
139	ダミー部材駆動機構	クランプ装置（44）				
140	無機質建材	耐火窓ガラス			○	
141	シリコーンエラストマー	エラストマーオルガノ ポリシロキサン		○		
142	平均粒子径 1～50μm の板状粉末	粒状フィラー		○		

143	常温で液状の油性成分	オイル		○		
144	脂肪族ポリエステル	ポリエステル PWD	○			
145	共重合ポリエステル樹脂	ポリエステル P1	○			
146	液状樹脂	エチレングリコール			○	
147	多価アルコール	ポリエチレングリコール			○	
148	縮合触媒	硬化促進剤			○	
149	研磨パッド用目詰まり防止剤組成物	研磨液組成物	○			
150	底板部	筒状部 21A				○
151	傾斜壁部	被覆部材 22				○
152	走行断続装置	走行クラッチレバー21	○			
153	変速手段	走行変速レバー18	○			
154	無段変速装置	摩擦円盤式無段変速機構	○			
155	固化材	注入材料液				
156	塊状体	グラウト流出部				
157	地盤改良工法	安定化処理方法				
158	注入手段	流出口				
159	地盤中	注入管周囲				
160	混合手段	混合室	○			
161	石炭灰	フライアッシュ		○		
162	一時貯留槽	バイパス管 3				
163	鋼管杭	中空杭	○			
164	固化材	セメントミルク			○	
165	走行体	走行台車 5	○			
166	アーム部	掘削アーム 11	○			
167	掘削部	カッタローダ		○		
168	掘削手段	掘削工具 13	○			
169	移送ベルト	ベルトコンベア 21	○			
170	移動可能	伸縮差動				

171	掘削土搬送手段	排土搬送装置 16	○			
172	回転体	旋回架台 4			○	
173	掘削装置	ケーソン工事用掘削機	○			
174	ポーラスコンクリートからなる鏡壁	ポーラスコンクリート部材 11	○			
175	控え壁	普通コンクリート部材 2	○			
176	連接壁	連結壁 10		○		
177	底壁	底 21	○			
178	筒状ブーム	ブーム 3	○			
179	筒型アーム	アーム 5	○			
180	２次圧縮空気	調圧空気	○			
181	渦流	螺旋状の空気渦流	○			
182	空気圧送用ブースタ	搬送ライン用流動化装置				
183	外管	出口管 20	○			
184	内管	入口管 10	○			
185	空気吹き出し間隙部	オリフィス 21				
186	横滑り防止ストッパ	固定ピン 8				
187	受枠	補強材 7				
188	ベース板	仕切板（9）	○			
189	補強材	剛性基枠部（13）				
190	床パネル	界床パネル 12	○			
191	耐火板	石膏ボード 30			○	
192	接着剤層	不陸調整剤	○			
193	連結具	パネル押え 2				○
194	パネル支持脚	パネル受部 1	○			
195	固定用フランジ部	円板 14		○		
196	通過孔	貫通孔 12	○			
197	ジャッキベース	台板 1			○	
198	円柱体	支柱 22a			○	

199	ジャッキハンドル	高さ調整用バー22b			○	
200	ベースプレート	建枠ベース 11	○			
201	鉄骨枠組	鉄骨躯体	○			
202	ハンドル本体	アウターハンドル 15	○			
203	ドア外板	アウターパネル 11		○		
204	ベース	フレーム 12				
205	台座	基体 1				○
206	プレート	可動片 2			○	
207	ヒンジ	蝶番 3		○		
208	アーム部	弾性係止板 4				
209	係合凹部	係止溝 4a，4b			○	
210	係合凸部	係止杆 5			○	
211	ガススプリング	ダンパー90			○	
212	取付片	ブラケット 91		○		
213	突出軸部	ボルト 86				○
214	固定具	割ピン 88		○		
215	ガススプリング	ステーダンパ 22			○	
216	突出軸部	ピン支持板 2				
217	固定具	ねじ軸 8			○	
218	プッシャ	開放部材 25				
219	スプリング	ばね部材 32		○		
220	押縁	ビード 5				
221	押縁コーナーブロック	コーナー部材 17	○			
222	ブロック取付部	対応辺 11				
223	第二引掛け部	拘止手段 5				○
224	第一引掛け部	係合部 6				○
225	フィルタ	煤捕集装置 19				○
226	窒素酸化物吸着ユニット	Nox 吸着剤 20		○		
227	三元触媒ユニット	触媒 54	○			

228	保護部材	レインフォースメント 8		○		
229	燃料噴射弁	インジェクタ 2		○		
230	制御手段	制御部 31	○			
231	蓄圧タンク	蓄圧器 52	○			
232	バルブボディ	弁端部 3			○	
233	噴孔	噴射開口 49		○		
234	噴孔プレート	孔付プレート 47	○			
235	負圧形成部	ガス供給フード 57				
236	負圧導入孔	ガス供給開口 71				
237	負圧形成部	エアアシストスリーブ 63				
238	噴孔プレート	オリフィスプレート 52	○			
239	環状壁部	案内部 65	○			
240	主体金具	取付金具（3）	○			
241	先端部	圧入面（37）				
242	プラグ配置孔	嵌着穴（63c）				
243	ポンプ	主ポンプ 1	○			
244	オートドレンバルブ	自動排出弁		○		
245	バルブ要素	弁体 13		○		
246	付勢部材	コイルスプリング 14			○	
247	流通路	貫通孔 28				
248	下流側連通路	排出系路 26				
249	上流側連通路	流入通路 18				
250	上流側開口部	上部本体 11				
251	オートドレンバルブ	自動水抜弁		○		
252	付勢部材	コイルばね 18		○		
253	下流側連通路	流出水路 10				
254	上流側連通路	流入水路 8				
255	表板	座金 36				○
256	裏板	アンカー対 31				○

257	引寄せ紐	連結ひも 35	○			
258	第 1 の取付部材	第一の取付金具	○			
259	第 2 の取付部材	第二の取付金具	○			
260	弾性本体部	本体ゴム	○			
261	ダイアフラム	支持ゴム				
262	仕切部材	オリフィス金具			○	
263	副液室	圧力制御室	○			
264	オリフィス通路	第一のオリフィス通路	○			
265	開口部	弁座面				
266	液封防振装置	液体封入式防振装置	○			
267	着座部	シールクリップ				
268	アクチュエータ	ソレノイドコイル			○	
269	ケース	ケーシング 1				
270	弾性部材	ブリーザ本体 13			○	
271	ハウジング	キャップ 14				
272	弾性部材	ルップ 24’				
273	ハウジング	カバー21			○	
274	シフトエレメント	シフトフォーク 12	○			
275	軸受エレメント（シフトシャフト）	シフトロッド 11	○			
276	シフトエレメント	シフトフォーク 12ad	○			
277	軸受エレメント（シフトシャフト）	シフトフォークシャフト 10	○			
278	かみ合いロック部材	ロックピン 34	○			
279	水室	シリンダ室 9	○			
280	水給排経路（水給排口）	連通孔 15				
281	第 2 のバネ	コイルスプリング 144		○		
282	流路部	オリフィス			○	
283	弁体受けテーパー部	テーパー部	○			

284	流入側テーパー部	逆テーパー部	○			
285	キャン	ケース				○
286	弁開閉口	オリフィス			○	
287	弁体受けテーパー部	円錐受座面			○	
288	流入側テーパー部	整流用円錐口			○	
289	開度維持板	板 30	○			
290	開度維持板	目盛板 8	○			
291	筒状のソケット	第 2 継手部材 17			○	
292	筒状のプラグ	第 1 継手部材 9			○	
293	チャック部材	コレット 29			○	
294	リングシール	シール部材 21	○			
295	パイロット燃料ノズル	燃料スプレーノズル 40	○			
296	アキシャルスワーラ	渦巻き翼 42,44		○		
297	ミキサ組立体	燃料噴射装置 10				○
298	燃焼器	ガスタービンエンジン燃料室 11	○			
299	加湿器	ヒータ式超音波加湿器	○			
300	殺菌チャンバ	ヒータ水槽 130			○	
301	熱交換機本体	サイロ			○	
302	空気供給装置	冷却空気送入配管	○			
303	空気排出装置	空気吸入配管	○			
304	外部入力パラメータ	独立運転条件データ				
305	内部生成パラメータ	従属運転条件データ				
306	自動設定手段	従属運転条件データ設定手段	○			
307	計量装置	組合せ秤			○	
308	制御手段	CPU4				
309	記憶手段	読み出し書込みメモリ 6				
310	固定側部材	土台 29				
311	可動側部材	皿 18				
312	ばね固定部	支持金具 27，金具 26			○	

313	複数枚の板ばね	板バネ 23	○			
314	固定側電極，及び，可動側電極	2 枚の板金電極 28	○			
315	有機樹脂	液状シリコン樹脂	○			
316	導電粒子	導電性粒子	○			
317	無機絶縁性微粒子	微粉シリカ			○	
318	導電粒子	導電性カーボンブラック	○			
319	無機絶縁性微粒子	無機質充填材			○	
320	荷重受け部	受圧部 2	○			
321	荷重受け部	荷重導入部 10	○			
322	荷重受け部	荷重印加部 1a	○			
323	センサベース	平板 22				○
324	操作端	力伝達部 23				○
325	第 1 の抵抗層	歪検知素子 24，25，27，28				○
326	第 2 の抵抗層	歪検知素子 26				○
327	突起電極	四角錐台部 9B				
328	ベース	円柱部 9A				
329	第 1 の金属層	チタン膜 25 及びタングステンカーバイドーコバルト合金			○	
330	第 2 の金属層	ニッケル合金			○	
331	板状基材	シリコン基板 21			○	
332	テスト回路	バイパス配線及びスイッチ				
333	テスト専用端子	入力パッド 21，40				
334	ウォータジェット方式の推進機構	ジェットポンプ			○	
335	音響測深器	音響測深器 23	○			
336	音響測深船	ジェットポンプ型推進船	○			
337	計測した水深データ	水深の測定データ	○			
338	送信手段	データ送信装置	○			
339	光検出素子	光電変換素子	○			

340	ホトダイオード	シリコンフォトダイオード	○			
341	トレンチ溝	分離用の溝	○			
342	メッシュ状補強体	トリコット編基布			○	
343	ホルダ	ボビン 12				
344	第 1 扁平コイル	駆動コイル 10c	○			
345	第 2 扁平コイル	駆動コイル 10d	○			
346	光学部材	偏光板			○	
347	振動検出手段	揺れセンサ 304 および 305		○		
348	カメラ	カメラ本体 301	○			
349	像揺れ補正手段	補正光学系 309	○			
350	像揺れ補正制御手段	CPU311				
351	レンズ	交換レンズ 308	○			
352	箱体	筐体		○		
353	光源ランプ	UHP ランプ 9	○			
354	画像形成部	DMD 表示デバイス 5				○
355	反射鏡	レンチキュラー11				○
356	投射結像装置	偏心フレネル 12				○
357	補正手段	スクリーン 1				○
358	リアプロジェクションテレビ	リアプロジェクション表示装置	○			
359	剥離層 5	再使用可能接着剤用剥離剤層 23			○	
360	突起部	作動部 3	○			
361	突起部	センサレバー47				○
362	画素電位側容量電極	保持容量用画素電極 2	○			
363	固定電位側容量電極	保持容量配線 4	○			
364	メール原稿	送信原稿	○			
365	電子データ	画像データ	○			
366	受信者のアドレス	受信者の宛先情報	○			
367	送信者の情報	送信元の情報	○			
368	送信装置	画像通信装置 601	○			

369	仲介管理装置	画像通信装置 602	○			
370	出力装置	コンピュータ端末			○	
371	ホスト装置	ワークステーション（WS）			○	
372	周辺装置	被制御装置（TS）	○			
373	インタフェースカード	ワークステーションアダプタ（WSADP）			○	
374	インタフェース制御部	MPU 部	○			
375	第 1 の接続コネクタ部, 第 2 の接続コネクタ部	2 つのネットワーク インタフェース部	○			
376	サービス利用者端末	クライアント端末	○			
377	サービス提供業者側管理サーバ	サーバ	○			
378	利用者識別子	ユーザ ID		○		
379	提携サイト	加盟店サーバ			○	
380	認証情報	入力された ID		○		
381	認証システム	認証サーバ	○			
382	表示媒体	LCD			○	
383	情報記録媒体	IC カード			○	
384	外部装置	カード R/W			○	
385	操作部	キースイッチ			○	
386	ID カード	IC カード	○			
387	2 枚	複数枚	○			
388	読取装置	IC カードリーダライタ			○	
389	利用資格データ	許可コード				○
390	車両識別コード	駐車アドレス				○
391	予約車両識別コード	出庫予約番号	○			
392	出庫情報表示装置	駐車出庫案内システム	○			
393	出庫識別コード表示手段	出庫位置案内表示器 5	○			
394	人体検知センサ	撮像手段				○
395	しきい値	所定値	○			

396	人工網膜センサ	人体検出センサ	○			
397	パッシブセンサ	赤外線センサ	○			
398	異常検知手段	異常判定手段 4	○			
399	異常検知情報	警報信号				○
400	排水栓	電磁弁				○
401	駆動手段	強制排水手段 6	○			
402	浴室システム	異常検出装置				○
403	監視装置	浴槽内事故検出装置	○			
404	監視情報	緊急情報	○			
405	通信網	電話回線			○	
406	異常検知手段	検出手段	○			
407	動作実行手段	強制排水手段	○			
408	動作完了状況	排出状態				
409	監視システム	浴槽内事故検出システム	○			
410	監視装置	検針メータ 100				
411	通信網	通信媒体 504	○			
412	動作実行手段	遮断弁 102				
413	制御手段	弁制御部 116	○			
414	緊急通報装置	情報通信端末			○	
415	温度管理室	冷蔵庫 2			○	
416	リーダ装置	運行管理計 5				○
417	車間距離センサ	距離検出手段	○			
418	ブレーキ加圧装置	走行制御装置	○			
419	制御部	制御手段	○			
420	シートベルトスイッチ	乗員状況センサ			○	
421	第 2 の磁性層	第一の磁気ヨーク層	○			
422	第 1 の磁性層	第二の磁気ヨーク層	○			
423	追加の磁気ヘッド	回転ヘッド 9，38	○			
424	2 値化手段	コンパレータ 3		○		

425	誤り検出手段	誤り訂正回路 4	○			
426	デジタル信号再生処理装置	情報記憶装置	○			
427	制御信号	内部データマスク信号 IDM	○			
428	DM レジスタ	マスク保持部 MASK				
429	DQ ライトドライバ	ライトアンプ	○			
430	プリント基板	回路基板（2）	○			
431	固定接点	リード線部（7a～9b）				
432	可動接点	接片（5）				
433	接触部	接触子（5a）	○			
434	係合孔	穴部（6a～6c）			○	
435	プリント基板	プリント基板（5A）	○			
436	固定接点	固定接片（6a）	○			
437	可動接点	可動接片（3）	○			
438	接触部	接点（32）		○		
439	スペーサシート	絶縁シート（4B）		○		
440	小孔	開口部（42L, 42R）			○	
441	通信手段	回線接続回路 20				○
442	データベース	課金データベース 25	○			
443	管理サーバ装置	サーバー26	○			
444	充電池	充電ターミナル 16	○			
445	中芯材	中心導体			○	
446	導体	外部導体	○			
447	カードコネクタ	コネクタ 26	○			
448	メモリカード	メモリカートリッジ 10	○			
449	送りローラ	ローラ 70	○			
450	押圧部材	引き込み爪 60				
451	表面ガラス	基板 1				○
452	有機 EL 積層膜	有機 EL 構造体 4	○			
453	背面ガラス	封止板 3				

5.3　評価実験と結果

本節では，前節までで述べた，抽出分類した複合名詞が実際に利用できるかどうかについて，類似評価した内容と結果を示し，考察について述べる．

以下は，特許庁のデータベースから複合名詞を取り出す手順を示す．

STEP1　データベースから IPC 分類 G06F の公開特許公報を選び出す
STEP2　選び出された公開特許公報の審査書類情報照会を行う
STEP3　特許査定，拒絶査定，それ以外にわける
STEP4　拒絶査定された出願についての拒絶理由通知書を確認
STEP5　拒絶理由通知書で特許法 29 条 1 項が理由になっているものを抽出
STEP6　STEP5 中で明確な対比箇所を抽出

複合名詞（単名詞も含む）を対比する際に，以下のルールにのっとる．

・不要な形態素などは除去する（接頭語，接尾語，数詞，記号など）．
・接頭詞は指示代名詞として機能する場合も多いが，もしこれらを除去しても意味的機能は失われないとする．

複合名詞の類似性評価

比較された複合名詞を評価し，実際に類似している複合名詞同士なのかどうかを判断する．

類似度の評価は評価 1～3 とし，評価 1 は類似～類似の可能性大とし，評価 2 については，類似の可能性は高いが，個々の文書を確認する必要があるものと定義する.評価 3 については,類似性はほとんどないものとして非類似とした.

本願から抽出した複合名詞 L_i(複合名詞 L は i 個の形態素で構成される)とする.

引用から抽出した複合名詞 M_j（複合名詞 M は j 個の形態素で構成される)とする.

N を,複合名詞を構成する形態素数とするとき,

i>j のとき，i=N とする

i<j のとき，j=N とする

図 5.7 を複合名詞判定ルールとして，複合名詞の類似判定を行い，判定分類する（図 5.8).

比較する複合名詞AあるいはBどちらかの形態素数(N)
・形態素数＝2の場合
Ⅰ 評価1 同義語数(類義語を含む)が2のとき 判定レベル:類似～類似の可能性大
Ⅱ 評価2 同義語数(類義語を含む)が1のとき 判定レベル:類似の可能性あり(人の目の判断が必要)
Ⅲ 評価3 それ以外 判定レベル:非類似

・形態素数＞2の場合
Ⅳ 評価1 同義語数(類義語を含む)がNのとき 判定レベル:類似～類似の可能性大
Ⅴ 評価2 同義語数(類義語を含む)が[N/2]のとき 判定レベル:類似の可能性あり(人の目の判断が必要)
Ⅵ 評価3 それ以外 判定レベル:非類似

例1） Ⅵより非類似判定
A:異常事象検知手段 N=4 （異常/事象/検知/手段）
B:画像比較部 N=3 （画像/比較/部）
例2） Ⅴより類似判定
A:処理割当装置 N=3 （処理/割当/装置）
B:処理管理装置 N=3 （処理/管理/装置）

図 5.7 複合名詞判定ルール

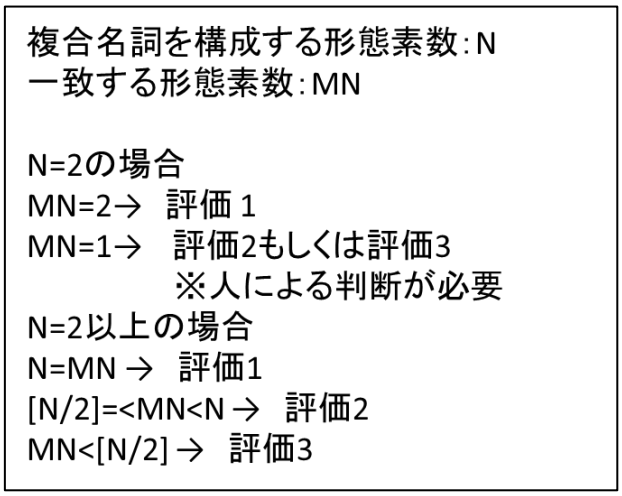


図 5.8　複合名詞判定分類

図 5.8 は，複合名詞を構成する名詞の類似数から複合名詞の類似を判断していった結果を分類するルールとしてまとめたものである．複合名詞の類似は，複合名詞を構成する形態素（名詞）の一致数が，構成の半数以上類似する場合，ほぼ類似しているという判断をする審査官の経験則に基づきルールを定めた．

対象については，セクション G に着目し，その中で出願件数のもっとも多い 06 クラスに絞った．G06 の出願の 9 割はサブクラス F(電子的デジタルデータ処理)であるので，G06F を分野付けしている公開特許公報を調査した．

対象データ
公開特許公報の期間：2006.01.01-2006.03.31
公開特許公報数：5,540 本
抽出複合名詞対：313 対
調査分野（IPC）: G06F

313 対のうち，文として抽出された 14 対は省いた．その結果，

評価 1：137 対　45.82%
評価 2：76 対　25.42%

評価 3：86 対　28.76%

という結果を得られた．再現率を上げるためには，評価 1 に加え，評価 2 を加える必要があると考えた．　評価 1 と評価 2 の合計を利用することで 70%を超え，類似複合名詞の辞書データとして利用できる可能性が高いと評価できる．

評価 1 と評価 2 の合わせたものを G06F データ用の同義複合語辞書データとした．

5.4　結言

本章で，拒絶理由通知書から抽出した複合名詞を辞書データとして扱うために，複合名詞の類似判定手法を提案し，複合名詞の類似度について検証した．1 つ目が，比較される複合名詞を含む文節をそれぞれに係り受け解析し，係り元，係り先の類似から，対象の複合名詞の類似を判定する手法，2 つ目が，比較される複合名詞自体を形態素解析し，複合名詞を構成する名詞から複合名詞の類似を判定する手法である．

まず，分野を特定しない検証において，1 つ目の係り受け解析を利用した手法と 2 つ目の形態素を利用した手法で複合名詞の類似を判断した．その結果，抽出した複合名詞は，すべてがそのまま辞書データとして利用できないが，高確率で辞書データとして機能させることができることがわかった．

次に，分野を限定した場合において，複合名詞の類似判定を行った．この結果，70%以上の抽出複合名詞が類似の可能性があるものとして，利用できることが明らかとなった．それゆえ，拒絶理由通知書から抽出した複合名詞のデータを辞書として利用することができることが示された．本章では，分野を限定した範囲内での検証であったが，他の分野においても有効であるかどうかについて，今後も検証していく．

しかし，ここで新たな問題点がわかった．4 章では，比較対象の文書について形態素の代わりに複合名詞を利用して文書間の類似度を測る有用性を示したが，文書中の複合名詞は，係り先の動詞によって異なる意味を持つ．例えば，複合名詞「情報処理装置」を「備える」のか，「移動させる」のかによって「情報処理装置」の持つ意味が変わってしまう．

よって，次章では，複合名詞のみを対象とせずに，複合名詞+動詞のフレーズを対象文書から抽出させ，作成した辞書を利用した文書間の類似度についての検証を行う．

第６章

複合名詞を利用した文書類似判定

本章では，4 章の複合名詞抽出方法，5 章の複合名詞の類似判定手法から複合名詞類似辞書データ（以下，辞書データ）を作成し，その辞書データを利用した文書類似の判定について述べる．まず，辞書データ作成について述べ，次に，類似判定対象データからフレーズ抽出と，作成辞書適用前後の文書類似度について述べる．その後，フレーズの集約と，それらを用いた文書類似判定について示す．

最後に，評価実験の内容と結果について述べ，本手法が特許文書の類似判定に有効であることを示す．

6.1　本願・引用の文書類似度比較

本願と引用の文書間の類似度の比較を行う.

類似判定対象データとして，G06F 分野（物理学 計算；計数 電気的デジタルデータ処理）を限定し，出願が公開されるときに発行される「公開特許公報」が 2006 年 1 月 1 日から 2006 年 3 月 31 日の期間に公開されたもの 5,540 本より，特許法第 29 条第 1 項 3 号 7)の要件を満たさずに最終的に拒絶査定となった出願．そのうち本願に対して引用箇所で実施例をあげている公報の請求項全文とその引用 140 本 70 組を用いた．そして，それぞれの請求項全文より，「複合名詞→動詞」の係り受け部分をフレーズとして抽出した．これらに 5.3 評価実験で利用した G06F データを適用させ，通常文書類似に利用するコサイン類似度と比較し，本願と引用の類似度の比較を行った．その際に，文書間類似度式を提案した.

6.1.1　利用辞書データ

本節では，文書類似に利用する辞書データについて述べる.

前章で，拒絶理由通知書から抽出した複合名詞を検索に利用可能な辞書データとして整備した辞書を，分野 G06F 辞書として，検証する際に利用する辞書データとして利用する.

6.1.2　フレーズ抽出

前章で述べたとおり，請求項の内容を的確に捉えるために，複合名詞だけでなく複合名詞を係り元として，係り先となる動詞とフレーズとして抽出する．

請求項からのフレーズの抽出には，数理研究所のテキストマイニングスタジオを利用した．図 6.1 は，抽出されたフレーズの例を示す．

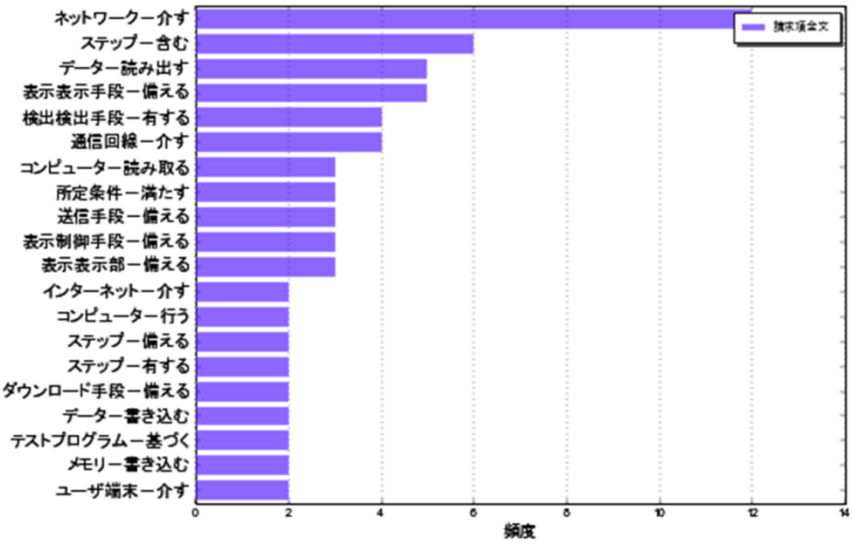


図 6.1　フレーズ抽出

そして，これらのフレーズの関係を，ノード（複合名詞と動詞）とその関係を表すリンクでネットワークする係り受け関係を整理したネットワーク図（図 6.2）を作成した．

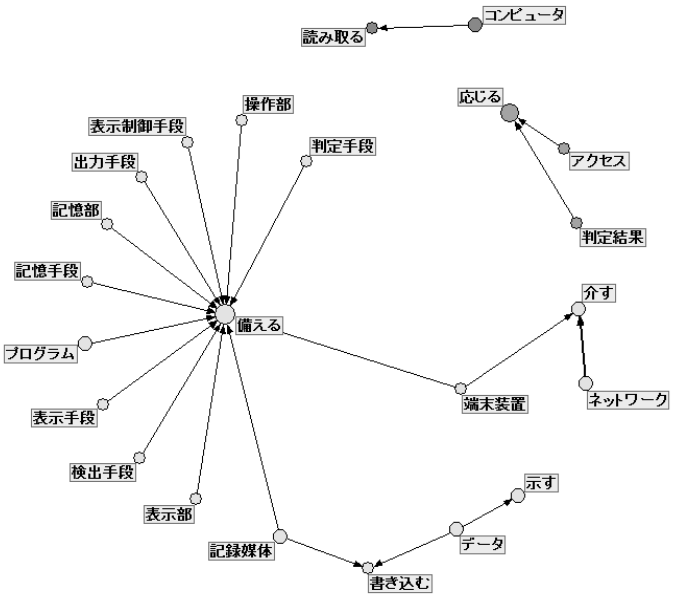


図 6.2　係り受けネットワーク

6.1.3　フレーズの比較

本願と引用の類似度を比較するために，文書間類似度式（式 6.1）を作成した．

$$F(sim)= \frac{1}{N}\sum_{i=1}^{N} k_i \alpha\beta \qquad (式\ 6.1)$$

式 6.1 中の α と β の計算を α の例（図 6.3）で示す．

複合名詞が係る動詞への寄与率を計算する．α は本願について，β は引用について，それぞれの複合名詞が動詞に係る寄与率を示す．

複合名詞 CN の係り先動詞を V とするとき，1 つの V に係る CN の数を Nx とすると寄与率は，式 6.2 で示される．

$$\alpha \text{ または } \beta = \frac{1}{Nx} \quad (式\ 6.2)$$

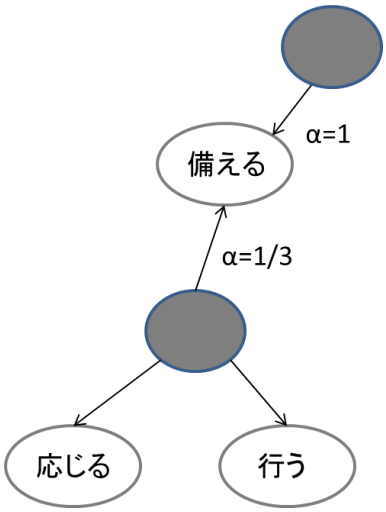


図 6.3 複合名詞の動詞への寄与率

N は本願のノード間のリンク数を表す．また，式 6.3 の Ki は係り先の動詞への類似度を計算したものであり，係り先である動詞が同じ場合は重みをそのまま 1 として計算し，違う場合には，ジャッカード係数を利用して定義した．

X，Y はそれぞれ本願と引用の係り先動詞の種類の数を示す．

$$K_i = 1 - \frac{|X \cap Y|}{|X \cup Y|} \quad (式\ 6.3)$$

ただし，本願リンク数を N 本，引用のリンク数を M 本としたときに，M の値が N の半分にも満たない場合は，補正を行った．なぜなら，M=1/2・N の場合，仮に引用からの抽出と本願からの抽出がすべて一致したとしても計算上類似度 50%が上限値となってしまう．このような抽出数による影響を少なくする

ため，一般的な加重平均ではなく，N と M の調和平均（式 6.4）とした.

$$\text{調和平均} = \frac{2MN}{M+N} \qquad \text{（式 6.4）}$$

6.1.4　フレーズによる類似度

表 6.1 は，実際に比較した 1 つを表記した.

表 6.1　本願と引用の係り受け類似解析

N= 11		cosine sim	27.472%		F(sim)=1/N(Σ kiα β)⇒	75.727%				
本願複合名詞	係り先 X=3	引用複合名詞	係り先 Y=5	本願	引用	X∩Y ＝ 2 sim=|X∩Y|/|X∪Y|= 0.333 K=1−sim=0.666	α		β	8.330
CPU指示	応じる	アクセス処理	応じる	CPU指示－応じる	プロセッサバス制御部－備える	0.666	1	1	0.666	
ノードメモリ	異にする	インタリーブ	応じる	ノードメモリー異にする	各ノード主記憶装置間－応じる	0.666	1	1	0.666	
ノード内	応じる	ノード	備える	ノード内－応じる	ノード－備える	0.666	1	1	0.666	
メモリインターリーブ	応じる	ノード間インタリーブ制御部	備える	メモリインターリーブ－応じる	インタリーブ－応じる	1	1	1	1	
異容量メモリ	備える	ノード番号	備える	異容量メモリー備える	分散共有メモリ装置－備える	1	1	1	1	
記録媒体	備える	ノード番号演算部	備える	記録媒体－備える	主記憶装置－備える	1	1	1	1	
制御チップセット	備える	プロセッサ	備える	制御チップセット－備える	プロセッサー備える	1	1	1	1	
設定宛先レジスタ	備える	プロセッサバス制御部	備える	設定宛先レジスター備える	レジスター備える	1	1	1	1	
転送速度	備える	レジスタ	備える	転送速度－備える					0	
複数ノード間	応じる	各ノード	備える	複数ノード間－応じる	ノード間インタリーブ制御部－備える	0.666	1	1	0.666	
複数設け	応じる	各ノード主記憶	割り当てる	複数設け－応じる	各ノード－備える	0.666	1	1	0.666	
		各ノード主記憶装置間	応じる							
		構成分散共有メモリ装置	備える							
		構成分散共有メモリ装置	用いる							
		主記憶装置	備える							
		主記憶容量	持つ							
		出力リクエストアドレス	応じる							
		分散共有メモリ装置	備える							
		保有情報	応じる							

この場合，コサイン類似度は 27.472%で，本手法の計算によると本願と引用の類似度は 75.727%となる.

図 6.4 では，本願で係り受け数が極端に少なく抽出できなかったものを除いた 58 組について全体の比較結果を示す.

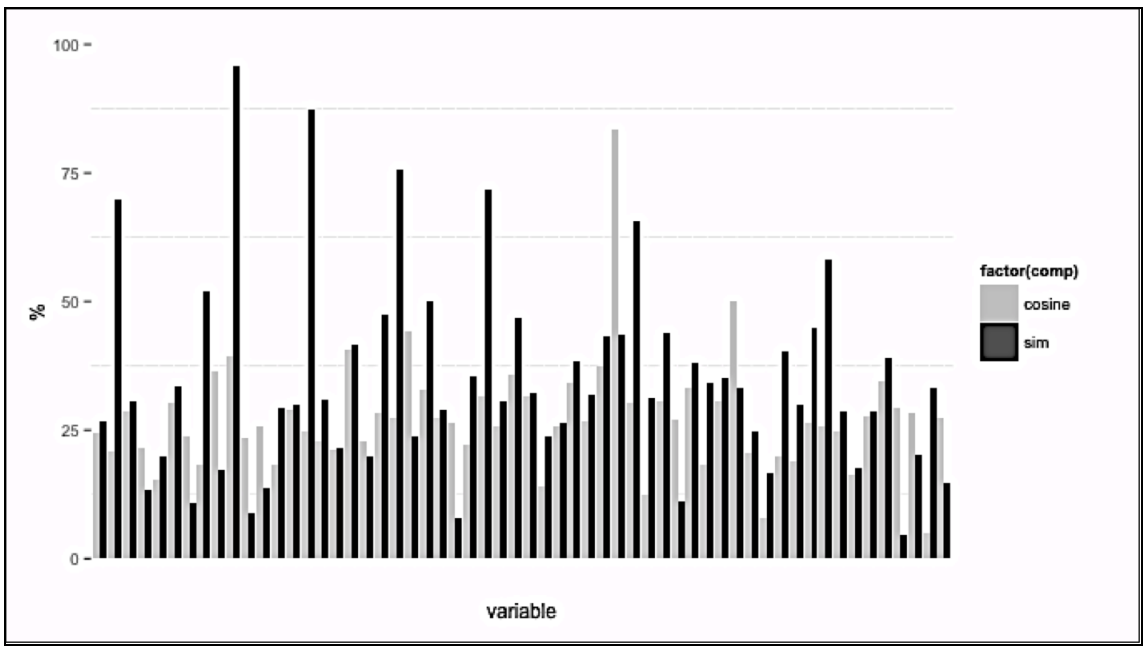


図 6.4　cosine 類似度と提案手法類似度の比較結果

6.1.5　本手法検証および考察

図 6.4 のうち 44 組で本手法の類似度が従来のコサイン類似度の手法を上回った．cosine 棒グラフがコサイン類似度を示したもので，sim 棒グラフが本手法での類似度算出をした結果である．これによると，類似である本願と引用のペアにおいて，本手法を用いた場合のほうが，従来の手法に比べ高い類似度を示すことができた．

コサイン類似度は，一致する形態素数を基に算出している．それゆえ，2 文書間で一致する形態素が多ければその文書間の類似度が高くなるという特徴がある．一方，提案手法は一致する複合名詞を基に文書間の類似度を算出する．そのため表 6.1 の例のように，一致する形態素の数は少ないが，形態素に分割される前の複合名詞の一致が多ければ，本手法での類似度が高くなるという結果が得られる．

コサイン類似度と本手法の類似度の相関係数の値 p=0.215476 と，相関が低

いことを示している．このことから，形態素を基にした場合と複合名詞を基にした場合では全く異なる類似度が算出されたこと言える．

表 6.2　コサイン類似度が低い 5 組のペア

cosine	sim
4.862	33.338
7.86	16.663
12.5	31.25
13.97	23.897
15.203	20

表 6.2 にコサイン類似度が低かった 5 組に対するコサイン類似度と本手法の類似度を示す．58 組のコサイン類似度の平均は約 27.4 である．このようにコサイン類似度が低い場合に，本提案では高い類似を算出している．本来，本願と引用は，特許審査官が類似と判断しており，高い類似度示すべき文書対であるため，コサイン類似度が低い事は課題であるが，本手法を用いることで，コサイン類似度だけでは不十分な部分を補完していると言える．

一方で，本手法の類似度が低くなる大きな原因は，形態素解析では同じ形態素が多く存在しても，複合名詞としての類似の判定が難しく，複合語が十分に抽出できないことに起因している．文書の中で十分な複合名詞の類似数がない場合の文書類似の判定については課題が残る結果となった．

また，2 つの類似度データについて，類似率を Wilcoxon の順位和検定を行った結果，$p<0.05$ となり有意水準 5%で平均値に差があるといえる．

得られた各データの比較結果，本手法を用いることにより，コサイン類似度が低かった文書対でも高い類似性レベルを得られ，類似性判断をより正確に行えるようになったといえる．

本提案手法を用いることにより，特許請求項に対する類似文書検索において，高い類似度を示すことがわかった．

形態素解析を利用し，コサイン類似度で文書の類似度を測る手法の場合，「情報記録用外部情報記憶装置」は「情報」「記録用」「外部」「情報」「記憶」「装置」という具合に名詞に分割されてしまい，本来抽出したい「情報記録用外部情報

記憶装置」のような特許文書独特の長い複合名詞に対して的確な類似度が測定できない．よって，複合名詞を利用しての類似度検索は有効であると 4 章および 5 章で述べたが，2 章で述べたように，特許の世界では，長すぎる複合名詞に対して，まったく同じ複合名詞を別の文書中で見つけることが困難であるという問題が生じている．

本研究では，分野を絞り，類似の正解として拒絶査定を受けた特許と引用された特許のペアで文書間類似度をテストすることにより，複合名詞を利用した類似度について指標を示すことに成功した．

6.2　複合名詞集約による文書類似度比較

抽出した複合名詞の辞書データさらに集約することで，文書の類似度の変化を検証する．類似度はフレーズをベクトルとしたコサイン類似度とする．
また，複合名詞の集約が正しくできているか判定をする．

6.2.1　使用辞書データの作成

下記の通り，分野を限定した辞書データを作成した．

対象 IPC：　G06F21/20
公開特許公報　2007.01.01-2007.12.31：　1,476 本
公開特許公報　2009.01.01-2009.12.31：　1,105 本
拒絶査定：　449 本
　　：　183 本
拒絶理由通知書からの対比抽出数：　359 組
　　：　276 組

抽出した合計 635 データを辞書データとして利用する．

6.2.2　辞書データ集約方法

STEP0 から STEP3 まで集約の方法を示す.

STEP0　検証対象データ
STEP1　抽出した辞書データ（1 項）を適用
STEP2　辞書データを word2vec で集約
STEP3　複合名詞内の形態素類似を利用した方法（5 章）で集約

複合名詞を集約していくことでより，図 6.5 の例のように，より上位概念の言葉にまとまっていくことになる.

この例では，最上位概念として,「認証処理」という言葉にまとまっていくまでを，STEP0 から STEP3 を通して集約していく.

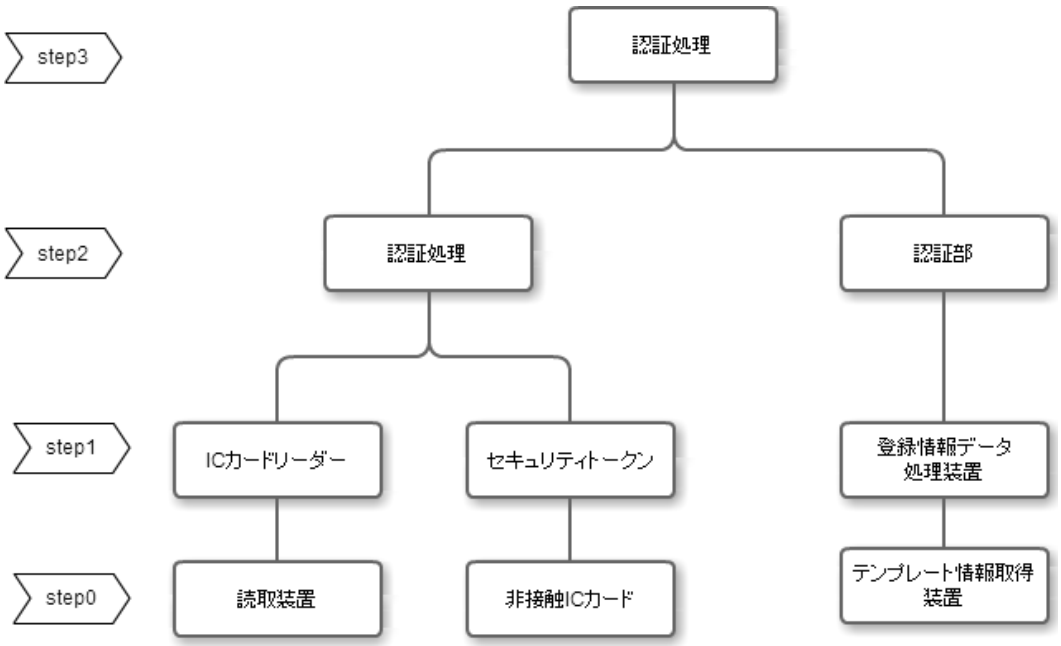


図 6.5　複合名詞集約

(1) STEP0

検証対象データそのままの状態で類似度を測る.

(2) STEP1

作成した辞書データを適用させて類似度を測る.

(3) STEP2

IPC 番号 G06F21/20 をもつ公開特許公報すべての 9,223 本から請求項を抽出し, 単語をベクトル化して表現する定量化手法である Skip-gram モデルを実装した word2vec [Mikolov 13] (https://code.google.com/p/word2vec/) を利用した.

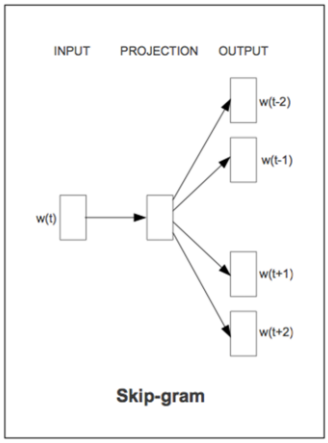


図 6.6　Skip-gram モデル

(Tomas Mikolov et. al. "Efficient Estimation of Word Representations in Vector Space")

Skip-gram は単語 wt が与えられて, 文脈中の 1 単語 wt+k を推定する問題である. 入力層に単語 wt を入れ正解データとして単語 wt+k を入れることを繰り返していく. 1 単語 wt の入力に対して, ニューラルネットワークの順方向の計算によって出力としてどの単語がどのくらいの確率で出現しそうか, という値が決まる. 次に, 正解である wt+k の確率が高くなるように, ニューラルネ

ットワークの重みを調整する．この繰り返しが学習であり，単語 wt のまわりの単語を予測できるように低次元埋め込み学習をする方法であるといえる．

ここでは，複合名詞と類似度の高い複合名詞を算出した．辞書データの複合名詞と類似度が高い複合名詞を図 6.7 の要領で集約した．

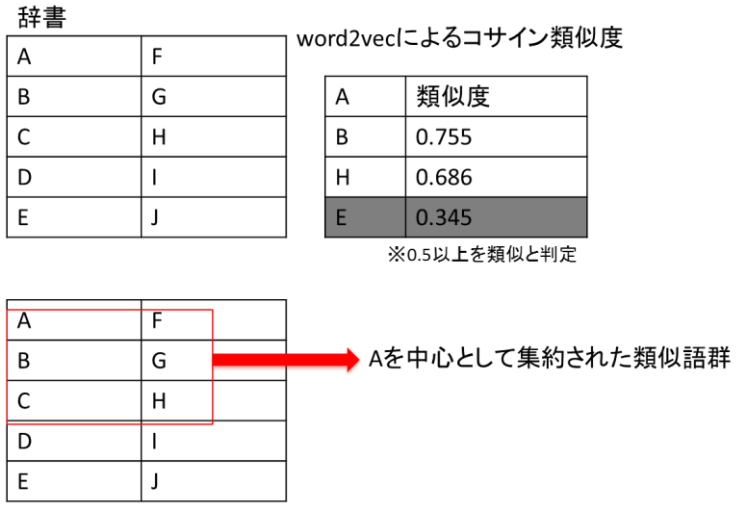


図 6.7　word2vec を利用した集約の考え方

(4) STEP3

5 章の図 5.8 のルールをもとに，可能である場合 STEP2 で集約した複合名詞をさらに集約する．

6.2.3　対象データ

発明の名称に「生体認証装置」という複合名詞をもち，IPC 番号が G06F21/20 をもつ公開特許公報のなかの請求項 1（メインクレーム）を抽出した．抽出数は，116 本とする．

データは，No.1 から No.116 まで 1 対 1 でそれぞれ文書間類似を比較する．類似比較総数は 6,670 対となる．

6.2.4　本手法検証および考察

STEP0からSTEP3までフレーズ集約をした結果を図6.8，表6.3に示した．これによると，フレーズ数が0.5倍に集約できると，文書間類似数は，5.3倍にあがることがわかった．

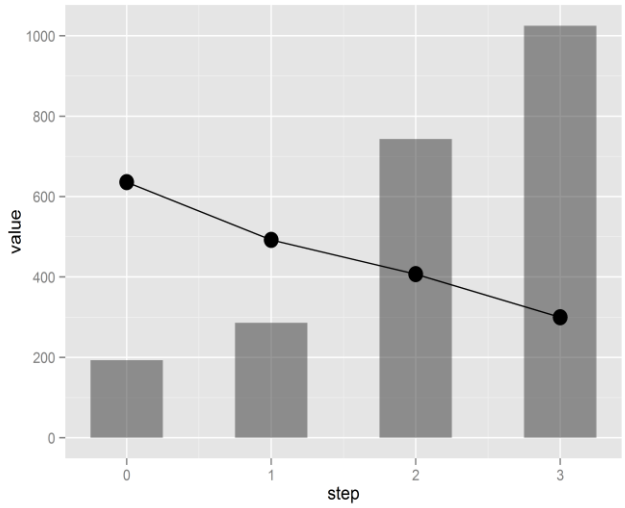


図 6.8　フレーズ集約数と類似文書数

表 6.3　フレーズ数と文書数

	フレーズの数	類似となった文書数
step0	635	193
step1	492	286
step2	407	743
step3	300	1025

また, データ 116 本の実際の内容は,「生体認証装置」の発明である 91 本と, それ以外に関する発明であるもの 25 本で構成されている．生体認証装置の発明である群を Y とし, 生体認証装置の発明ではない群を N とするとき, データの変数を文書から得られる複合名詞 CN とする場合，目的変数 Z で判別される（式 6.5）により，2 つの群を判別する（α は，Z を得るための定数とする）.

$$Z=\alpha_1 CN_1+\alpha_2 CN_2+\ldots+\alpha_x CN_x+\alpha_0 \quad \text{（式 6.5）}$$

線形判別分析を利用し，集約前と後それぞれにおける群の判別を行うことでその集約の方向性が正しいものかどうかを判定した（表 6.4，図 6.9）.

表 6.4　判別分析によるエラー判定

	n	y	エラー判定	n	y	エラー判定	
n	12	13	0.520	22	3	0.120	25
y	2	89	0.022	2	89	0.022	91
							116

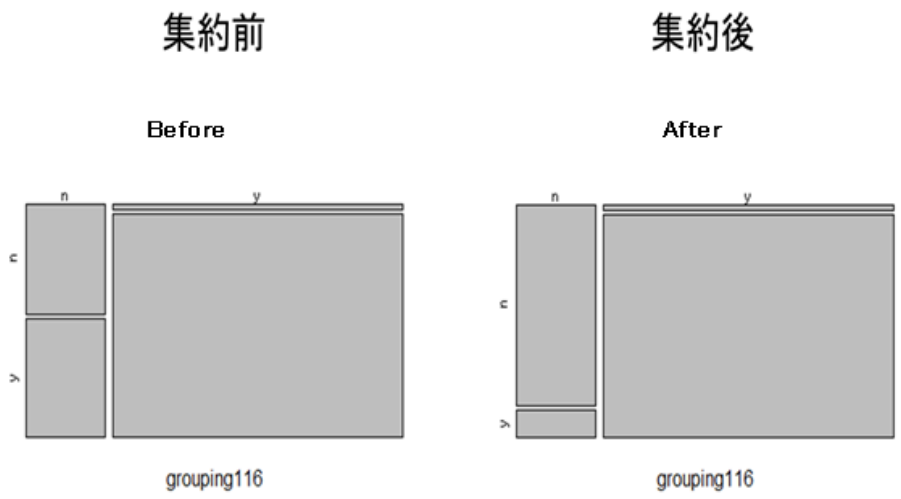


図 6.9　集約前後の分類判定

これによると，「生体認証装置」の発明ではないが，判定で「生体認証装置」の発明であると分類されてしまったデータが，集約前と集約後で 13 本から 3 本へ減ったことで集約が正しい方向で行われたことが示された．

また，比較した文書同士が「生体認証装置」の発明であるものを正解データとすることと定義すると，全正解データ数は 4,095 になる．

これを踏まえ，集約前の STEP0(Before)と集約後の STEP3 まで(After)の文書間類似のうち，正解データと合致する割合を，適合率，再現率，それらの調和平均値である F 値の数値を表 6.5 で確認する．

表 6.5　適合率・再現率・F 値

	Before	After
Precision	0.819	0.712
Recall	0.039	0.178
F-measure	0.074	0.285

Before（集約前）での類似数は 193 であり，その 193 の類似数のうち正解のデータ数は 158 である．一方，After（集約後）での類似数は 1025 であり，その 1025 の類似数のうち正解データ数は 730 である．

これらをまとめると，「生体認証装置」について類似の文書について，類似する文書間類似数は 4,095 である．

そのうち，Before および After の比較において，再現率が 0.039 から 0.178

に上昇したことがわかる．そして，このときの F 値は，0.074 から 0.285 に改善した．

これにより，複合名詞を集約することで，本来類似であるとされるべき文書において類似判定ができるようになったことが示される．

6.3　結言

6.1 では，本願と引用の比較対象文書から複合名詞と動詞のフレーズを抽出し，辞書データを適用させた文書の類似度と，従来の手法である形態素解析から算出したコサイン類似度を比較した．コサイン類似度で類似度が低くなってしまう文書についても，本手法では高い類似度を示すことができた．

また，6.2 では，作成した辞書データを利用し複合名詞を集約させた．比較対象文書の類似文書数は，複合名詞を集約していくにつれて多くなることが示された．また複合名詞が正しく集約できたかについて，判別分析を利用して確認することができた．

ここでは，まず辞書データから出現した複合名詞を上位概念をもつ複合名詞へ集約していき，それをさらに集約することで，実質的に類似複合名詞の数が増やせたとみなすことで，類似する文書数が増えたと考えられる．

これらの結果より，拒絶理由通知書から複合名詞を抽出し，辞書データ化して文書類似判定に利用する本手法は，有効であることが示された．

第 7 章

結論

本論文では，特許文書の類似度判定の精度向上を目的に，特許審査の過程で示される拒絶理由通知書を利用し，その中から複合名詞対を抽出し，類似複合名詞の辞書データを作成し，文書類似度判定に適用させる手法を提案した．

従来の自然言語処理技術による形態素解析を利用した方法では，特許文書のような特殊文書においては十分な精度を出すことが困難であった．そのため，特許文書の検索においては無関係な文書が類似と判定されるような結果も示されていた．本研究では，これらの問題を解決するために，複合名詞を形態素解析せず，複合名詞のまま利用することで，文書の類似度を測ることを目的とした．そして，類似複合名詞を同定することができる辞書データを作成し，この辞書データをもとに，複合名詞と動詞のフレーズを用いた文書間類似度，フレーズを集約することで，文書間類似度を高める手法を提案した．

本研究を行う動機は，著者が修士学生時代に知的財産学全般を学ぶなかで，難読な特許文書の理解に苦労したこと，文書の大分類は機械化で処理できる作業であっても，最終分類のところでは人手に頼らざるを得ず，そのうえ読み手の経験や勘所で判断されている現実を目の当たりにしたこと，そして，難しいといわれる特許文書を理解するためには，どう理解すべきか，どのような文書同士が似ていると判断されるべきなのか，そのための判断材料に何を用いるべきかなどの問題を意識したことである．

本論では，特許文書の難読の原因の 1 つでもある頻出する複合名詞に着目した．特許文書では，大部分の文書が類似になってしまう形態素解析を用いた類似度判定方法の問題点を解決するために，複合名詞のままの形で類似文書検索

に利用する方法を提案した.

しかしながら，複合名詞を利用すれば文書が何について書かれているか判断できる反面，文書中に同じ複合名詞が出現する率が少ないという問題に直面した.

この課題に取り組むために，類似する複合名詞の辞書の利用が重要であると考えた. しかしながら, 何をもって辞書データとするかを考える必要があった. 既存の類語辞書は, 新語が加わるまでにタイムラグが生じるという課題がある. そこで，著者は先行研究を行い，その過程で，特許の出願書類に関して様々な解析や分析が行われてきたが，審査官が示す審査書類については，ほぼ研究がなされていないことに気づいた．これにより，審査書類を段階的に確認していくことで，審査官が示す拒絶理由通知書の中に対比箇所があることを発見したことが本研究の核となった．そして，この記載部分を辞書データとして利用するという判断より，本研究に至った.

本論文の第 2 章および第 3 章では，通常の文書と特許文書との特徴の違いからどのように特許文書を分類していくか，特許文書中に出現する複合名詞についての説明と，特許制度について述べた.

第 4 章では，複合名詞を利用した文書検索について，形態素解析を用いた手法の結果を受けて，問題点を明らかにした．形態素解析を利用した文書類似検索の結果では，再現率は高いが，多くの非類似文書も類似となってしまい，効率的な検索ではないことを示した．一方，複合名詞解析では，複合名詞の出現率が低くなるため，再現率を高める必要がある．この課題に対して，同じ意味を持つ複合名詞同士を類似複合名詞として辞書データ化する必要性について論じた.

第 5 章では，拒絶理由として特許法第 29 条第 1 項 3 号の新規性の欠如を理由とした拒絶理由通知書から，審査官が対比して類似と判断している複合名詞対を抽出した．複合名詞の類似度評価では，文字列照合，同義語･類義語，シソーラス，係り受け手法などから抽出したデータのうち，80.8%で辞書として利用可能な抽出であったことを示した. そして, 実際の検索に適用できるように, IPC 番号 G06F について対象分野を絞った複合名詞抽出に取り組んだ．その結果，類似する，あるいは類似する可能性があると判定できる複合名詞対の割合

が 70%強となった．これにより，拒絶理由通知書からの辞書データの作成することが可能であることを確認した．

本論文の第 6 章では，まず第 5 章で作成した辞書データを用い，本願，引用それぞれの特許文書から複合名詞と動詞のフレーズを抽出したものに適用させた．その際の類似度の指標に文書間類似度式を作成し類似度を算出した．

また，通常の文書比較で用いられる形態素解析をベースとしたコサイン類似度を用いた文書間類似度との比較検討を行い，コサイン類似度で低い類似度になってしまった文書間類似度について，本手法では高い類似度を示すことができたため，コサイン類似度だけでは不十分な部分を補完することができることを確認した．

さらに分野を絞り込み，新たな辞書データを作成した．辞書データを構成する複合名詞と動詞のフレーズを集約していくことにより，比較する文書間類似数が増加した．フレーズ集約が 0.5 倍になると，文書類似数が 5.3 倍になるという結果が得られた．集約したフレーズについても，線形判別分析にて集約が正しい方向で行われていることを示した．

以上，本研究で提案する「拒絶理由通知書情報を活用した特許文書類似度判定に関する研究」について報告した．

拒絶理由通知書から複合名詞を抽出する利点は，審査は特許庁で常に行われており，一定の数は拒絶査定を受け，その際に拒絶理由通知書が必ず示される．この拒絶理由通知書の利用の有益な点は，類似の判断を特許審査官が行っているという点にある．本来，類似の判断は最も難しいが，その判断を特許審査官が担うことで，類似の点で高い信頼性をもつ辞書データを得ることができる．また，この辞書データは，特許審査が行われる限り自動的に蓄積することができる点でも非常に有益といえよう．

そのうえ，利用されている複合名詞は，時代とともに陳腐化される恐れがあるが，その時代ごと蓄積されてきた拒絶理由通知書のデータにおいて，複合名詞が使われていた時代背景をくみ取り，ある時代にだけ使われていた複合名詞だとしても，その複合名詞に対して類似とされていた別の複合名詞を判定することができる．

上記利点より，これまで解析されてきた出願書類ではなく，いままで目をつ

けてこられなかった審査書類の 1 つである拒絶理由通知書を利用する利点は大きいものであるといえる.

このような利点をさらに生かし，今後は大規模な辞書化手法の構築へと発展させることを考えている.

謝辞

博士課程の在学中，公私にわたり，大変お世話になった筑波大学大学院 ビジネス科学研究科の 津田 和彦 教授に深く感謝いたします.

副指導教官をお引き受け頂き，的確なアドバイスを頂戴した筑波大学大学院ビジネス科学研究科の吉田 健一 教授，倉橋 節也 准教授に深く感謝いたします.

著者が大学院修士課程よりお世話になり，博士課程進学へ強く勧めてくださった東京理科大学専門職大学院イノベーション研究科の 平塚 三好 教授に深く感謝いたします．また，特許文書解析の道へすすむきっかけを与えてくださった「特許工学」の著者であり，IRD 国際特許事務所所長 弁理士 谷川 英和 先生に深く感謝いたします.

発表会の場などで，様々なアドバイスやコメントを下さった筑波大学大学院ビジネス科学研究科の教官の方々に，深く感謝いたします．また，研究の遂行にあたり，ご助言をいただきました東京理科大学専門職大学院イノベーション研究科の先生がたに深く感謝いたします.

津田研究室の方々には，日頃より研究の進め方についての貴重な示唆やご意見を頂戴いたしました．また，ON と OFF のメリハリの効いたゼミのメンバーと出会えたことは，私にとって貴重な財産となっております．メンバーの皆様に対し，ここに深く感謝いたします.

また，在学中に結婚することになりましたが，自身の仕事が忙しいにもかかわらず，文句１つ言わず，家事や私の精神面のサポートをしてくれ，疲れたときには笑わせてくれ，学会遠征にも同行してくれた主人，椛澤 英康に深く感謝いたします.

そして，わたしの妹ファミリー，大好きな姪っ子と甥っ子，家族の存在が何としても研究を完遂しようと思うわたしの最大のモチベーションとなりました．皆の存在と励ましに深く感謝いたします.

最後に，いままで親孝行と呼べることは何一つせず，大学時代からダンスばかりやってきたうえに，ダンス競技を引退してからも落ち着くどころか今度は大学院で勉強するという私に，半ば呆れながらも見守ってくれたわたしの両親，柳堀利雄と柳堀登志子に娘から最大の感謝を述べ，この研究の成果を捧げ，深く感謝いたします．

お世話になりました皆さま，本当にありがとうございました．

参考文献

[Ando 07]

安藤俊幸: “特許情報の分析・評価支援”, 情報プロフェッショナルシンポジウム予稿集, pp.13-17(2007)

[Ando 09]

安藤俊幸: “テキストマイニングと統計解析言語 R による特許情報の可視化”, 情報管理 52 , pp.20-31(2009)

[Asahara 03a]

浅原正幸, 松本裕治: “形態素解析システム「茶筌」version 2.3.3 使用説明書”, 奈良先端科学技術大学院大学 (2003)

[Asahara 03b]

浅原正幸, 松本裕治: “ipadic version 2.7.0 ユーザーズマニュアル”, 奈良先端科学技術大学院大学 (2003)

[Baba 15]

馬場錬成: “特許侵害判決の賠償額が低すぎる日本の裁判”, ㈱発明通信社コラム, 潮流 No.60 http://www.hatsumei.co.jp/column/detail/2/60.html（2015.04.16 参照）

[Baba 98]

馬場 肇: “日本語全文検索システムの構築と活用”, ソフトバンク(1998)

[Brown 92]

P.F. Brown, P.V. Desouza, R.L. Mercer, V.J.D. Pietra, J.C. Lai: “Class-based n-gram models of natural language”, Computational linguistics 18, pp.467-479(1992)

[Cai 04]

L. Cai, T. Hofmann: “Hierarchical document categorization with support vector machines”, Proceedings of the thirteenth ACM international conference on Information and knowledge management, pp. 78-87(2004)

[Chu 08]

X.-L. Chu, C. Ma, J. Li, B.-L. Lu, M. Utiyama, H. Isahara: “Large-scale patent classification with min-max modular support vector machines”, Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, IEEE, pp. 3973-3980 (2008)

[Deguchi 06]

出口昌信, 岡本和彦: “宇部興産（株）におけるエンドユーザー教育―営業部門に対する知的財産情報教育の試み―”, 情報知識学会誌 16 , pp.39-43 (2006)

[Eguchi 04]

江口浩二: “Web 検索の技術動向と評価手法”, 情報処理 Vol.45 No.6 (2004)

[Fall 03]

C.J. Fall, A. Törcsvári, K. Benzineb, G. Karetka: “Automated categorization in the international patent classification”, ACM SIGIR Forum, ACM2003, pp. 10-25 (2003)

[Fujii 02]

藤井敦, 石川徹也: “World Wide Web を用いた事典知識情報の抽出と組織化”, 電子情報通信学会論文誌, Vol.J85-D-II, No.2, pp.300 (2002)

[Fujii 08]

藤井敦: “特許情報を用いた辞典検索システム”, 情報処理学会研究報告, pp.9-15 (2008)

[Hanai 05]

花井拓也, 山村毅: “単語間の依存性を考慮したナイーブベイズ法によるテキスト分類 (類似性の発見)”, 情報処理学会研究報告. 自然言語処理研究会報告 2005 , pp.101-106 (2005)

[Hashimoto 08]

橋本力, 河原大輔, 吉田節行, 後藤広樹, 横山晶一: “特許文書の構文解析”, 自動制御連合講演会講演論文集 51 , pp.200-200 (2008)

[Hashimoto 12]

橋本泰一, 藤井敦: “特許文書のための形態素解析辞書の構築”, 言語処理学会第 18 回年次大会発表論文集, pp.789-792 (2012)

[Hearst 93]

M.A. Hearst, C. Plaunt: “Subtopic structuring for full-length document access”, Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 59-68 (1993)

[Hearst 94]

M.A. Hearst: “Multi-paragraph segmentation of expository text”, Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 9-16 (1994)

[Hearst 98]

M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, B. Scholkopf: “Support vector machines”, Intelligent Systems and their Applications, IEEE 13, pp.18-28 (1998)

[Hirano 05]

平野耕一, 古林紀哉, 高橋淳一: “日本語圏ブログの自動分類（分類, ブログ)”, 情報処理学会研究報告. 自然言語処理研究会報告 2005, pp.21-26 (2005)

[Hiratsuka 09]

平塚三好, 大澤紘一: “情報通信・エレクトロニクス産業の発展を阻害するパテントトロールへの対応策: 米国の懈怠の法理を導入する試み”, 日本経営工学会論文誌 60 , pp.145-152 (2009)

[Hiratsuka 13]

平塚三好: “IT・ソフトウェア特許の新潮流~ 活用・防御から標準化まで~: 5. ソフトウェア産業の発展を阻害するパテントトロールへの対策”, 情報処理 54 , pp.215-219 (2013)

[Ichikawa 01]

市川伸治: “ビジネスモデル特許検索にも有効な概念検索”, 知的資産創造 9,

pp.10-13 (2001)

[Inoue 01]

井ノ上直己, 帆足啓一郎, 橋本和夫: “文書自動分類手法を用いた有害情報フィルタリングソフトの開発”, 電子情報通信学会論文誌 D84 ,pp.1158-1166 (2001)

[Inui 06]

乾健太郎,浅原正幸: “自然言語処理の再挑戦: 統計的言語処理を超えて (< 特集> テキストの可視化と要約)” ,知能と情報, 日本知能情報ファジィ学会誌 18, pp669-681 (2006)

[Ishida 04]

石田由利子: “エンドユーザー教育と検索システムの選択 (< 特集> 特許検索に必要なスキルと知識)”, 情報の科学と技術 54, pp.240-247 (2004)

[Iwayama 03]

M. Iwayama, A. Fujii, N. Kando, Y. Marukawa: “An empirical study on retrieval models for different document genres: patents and newspaper articles”, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 251-258 (2003)

[Izumi 10]

和泉潔, 後藤卓, 松井藤五郎: “テキスト情報による金融市場変動の要因分析”, 人工知能学会論文誌 25, pp.383-387 (2010)

[J-PlatPat 15]

特許情報プラットフォーム

http://www.inpit.go.jp/j-platpat_info/reference/index.html (2015.04.01 参照)

[JPO 15]

特許庁

https://www.jpo.go.jp/indexj.htm (2015.03.15 参照)

[JUMAN 12]

日本語形態素解析システム

http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN (2015.04.01 参照)

[Kato06]

加藤亮, 橋本博之, 辻河登: “特許情報解析システム (第一報)”, 情報プロフェッショナルシンポジウム予稿集,pp. 5-9 (2006)

[Kawahara 06]

D. Kawahara, S. Kurohashi: “A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis”, Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Association for Computational Linguistics 2006, pp. 176-183 (2006)

[Kimura 05]

木村昌臣, 古川裕之, 塚本均, 田崎久夫, 空閑正浩, 大倉典子, 土屋文人: “医薬品使用の安全性に関するアンケートの解析 テキストマイニング手法の適用”, 人間工学 41 pp.297-305 (2005)

[Kiriyama 06]

桐山勉, 長谷川正好, 川島順, 大山勝弘, 都築泉, 玉置研一, 田中宣郎, 藤嶋進: “特許情報のテキストマイニングの可視化”, 情報プロフェッショナルシンポジウム予稿集 2006 , pp.11-15 (2006)

[Kita 99]

北研二: “確率的言語モデル”, 東京大学出版会, (1999)

[Kita 02]

北研二, 津田和彦, 獅子堀正幹: “情報検索アルゴリズム”, 共立出版 (2002)

[Kitahara 01]

日本国語大辞典（第二版）（小学館 2001 ）北原保雄 小学館国語辞典編集部

[Kobayashi 96]

小林義行, 徳永健伸, 田中穂積: “名詞間の意味的共起情報を用いた複合名詞の解析”, 自然言語処理 3 , pp.29-43 (1996)

[Kokubu 99]

国分芳宏:” 記事洪水の方舟ー全文検索システム” 情報管理 42(5), pp.380-389 (1999)

[Koller 97]

D. Koller, M. Sahami: “Hierarchically classifying documents using very few words”, (1997)

[Kondo 07]

近藤陽介, 佐藤理史: “多項ナイーブベイズ分類を用いた日本語テキストの難易度判定手法の検討”, 言語処理学会第 13 回年次大会発表論文集, pp.534-537 (2007)

[Konishi 04]

小西一也, 北内啓, 高木徹: “発明の特徴に着目した検索語抽出による先願特許検索”, DEWS2004, (2004).

[Konishi 06]

小西一也, 新海正吾, 高木徹, 三部靖夫: “先行技術調査を効率化する請求項理解支援機能”, 情報処理学会研究報告. EIP,[電子化知的財産・社会基盤] ,pp.19-26 (2006)

[Koyano 07]

古屋野浩志:” 特許分類等の付与精度向上への取り組み”, Japio 2007 YEAR BOOK, pp.118-119 (2007)

[Kudo 01]

工藤拓: 形態素解析器 MeCab, (2001)

[Kudo 05]

工藤拓: “形態素周辺確率を用いた分かち書きの一般化とその応用”, 言語処理学会第 11 回年次大会発表論文集, pp.592-595 (2005)

[Kumamoto 99]

熊本睦, 島田茂夫, 加藤恒昭:” 概念ベースの情報検索への適用: 概念ベースを用いた検索の特性評価”, 電子情報通信学会技術研究報告. 人工知能と知識処理 98 , pp.9-16 (1999)

[Kunishi 11]

国司洋介: “特許の可視化と特許解析 解析ツールとの付き合い方を考える”,

情報管理 53 , pp.591-599 (2011)

[Kurita 03]

栗田多喜夫: “サポートベクターマシン入門”, 産業総合技術研究所 http://www.neurosci.aist.go.jp/~ kurita/lecture/svm, (2003) (2015.03.01 参照)

[Mase 05]

H. Mase, T. Matsubayashi, Y. Ogawa, M. Iwayama, T. Oshio: “Proposal of two-stage patent retrieval method considering the claim structure”, ACM Transactions on Asian Language Information Processing (TALIP), 4, pp.190-206 (2005)

[Mase 07]

H. Mase, M. Iwayama: “NTCIR-5 patent retrieval experiments at Hitachi”, Proceedings of the 6th NTCIR workshop meeting, pp. 403-406 (2007)

[Matsumoto 00]

松本裕治: “形態素解析システム 「茶筌」(< 特集> 使いやすくなった自然言語処理のフリーソフト: 知っておきたいツールの中身)”, 情報処理 41 , pp.1208-1214 (2000)

[Matsumoto 03]

松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸: “形態素解析システム 『茶筌』 version 2.3. 3 使用説明書”, 奈良先端科学技術大学院大学 情報処理科学研究科 自然言語処理学講座, (2003)

[McCallum 98a]

A. McCallum, K. Nigam: “A comparison of event models for naive bayes text classification”, AAAI-98 workshop on learning for text categorization, Citeseer, pp. 41-48 (1998)

[McCallum 98b]

A. McCallum, R. Rosenfeld, T.M. Mitchell, A.Y. Ng: “Improving Text Classification by Shrinkage in a Hierarchy of Classes”, ICML, pp. 359-367 (1998)

[Mikolov 13]

T. Mikolov, I. Sutskever, K.Chen, G.S. Corrado, J. Dean: “Distributed Representations of Words and Phrases and their Compositionality”,Neural Information Processing Systems (NIPS), pp.1-9 (2013)

https://code.google.com/p/word2vec/

[Muguruma 01]

六車正道: “特許情報検索の課題と概念検索システムの役割”, 知財管理 51 ,pp.1891-1900 (2001)

[Muguruma 03a]

六車正道: “概念検索による特許情報の活用法 (1)”, 発明 100, pp.48-53 (2003)

[Muguruma 03b]

六車正道: “概念検索による特許情報の活用法 (2)”, 発明 100 , pp.52-56 (2003)

[Murakami 06]

村上明子, 渡辺日出雄: “時系列情報を利用した複合語キーワードの抽出(抽出, コーパス)”, 情報処理学会研究報告. 自然言語処理研究会報告 2006 , pp.1-8 (2006)

[Murawaki 10]

村脇有吾, 黒橋禎夫: “形態論的制約を用いたオンライン未知語獲得”, 自然言語処理 17 ,pp.55-75 (2010)

[Muto 00]

武藤晃: “特許情報の難しさの周辺”, 情報管理 43, pp.604-614 (2000)

[Nagao 78]

長尾真, 辻井潤一, 山上明, 建部周二: “国語辞書の記憶と日本語文の自動分割”, 情報処理 19 , pp.514-521 (1978)

[Nakagawa 03]

中川裕志, 湯本紘彰, 森辰則, “出現頻度と連接頻度に基づく専門用語抽出”, 自然言語処理 10, pp.27-45 (2003)

[Nakamura 05]

類語新辞典（三省堂 2005 ）中村 明 (編集), 森田 良行 (編集), 芳賀 綏 (編集)

[Nanba 05]

難波英嗣: “論文間の引用情報を利用した関連用語の自動収集”, 言語処理学会 第 11 回年次大会, (2005)

[Nanba 07]

H. Nanba: “Query expansion using an automatically constructed thesaurus”, Proceedings of the 6th NTCIR Workshop, pp. 414-419 (2007)

[Nanba 09]

難波英嗣, 釜屋英昭, 竹澤寿幸, 奥村学, 新森昭宏, 谷川英和: “論文用語の特許用語への自動変換”, 情報処理学会論文誌. データベース 2 , pp.81-92 (2009)

[Nanba 11]

難波英嗣, 竹澤寿幸, 乾孝司, 岩山真, 橋田浩一, 橋本泰一, 藤井敦: “特許検索履歴を用いたシソーラスの自動構築”, 言語処理学会第 17 回年次大会, pp.900-903 (2011)

[Nasukawa 09]

那須川哲哉: “テキストマイニングとは何か, テキストマイニングを使う技術/作る技術”, pp.1-64 (2009)

[Oinuma 13]

生沼貴大, 天沼 博, 松澤和光 : ” 文章間の類似度計算システム”,情報処理学会第 75 回全国大会（2013）

[Okabe 06]

岡部貴博, 吉川大弘, 古橋武: “インシデントレポート解析のための多重接続型階層的テキストマイニング手法の提案”, 日本知能情報ファジィ学会 ファジィ システム シンポジウム講演論文集 22, pp.54-54 (2006)

[Okamoto 06]

岡本和彦, 出口昌信: “山口 TLO に対する知的財産情報教育支援: 産学連携における試み”, 情報知識学会誌 16 , pp.44-51 (2006)

[Okumura 12]
奥村学, 藤井敦, 谷川英和, 岩山真, 難波英嗣, 山本幹雄, 内山将夫, “特許情報処理 : 言語処理的アプローチ”, コロナ社, (2012)
[Okutsu 75]
奥津敬一郎: “複合名詞の生成文法 国語学”, 東京明治書院, (1975)
[Oono 81]
角川類語新辞典(角川書店 1981) 大野 晋 (著), 浜西 正人 (著)
[Saito 92]
斎藤倫明: “現代日本語の語構成論的研究: 語における形と意味”, ひつじ書房, (1992)
[Sakachi 10]
坂地泰紀, 野中尋史, 酒井浩之, 増山繁: “Cross-Bootstrapping: 特許文書からの課題・効果表現対の自動抽出手法”, 電子情報通信学会論文誌 D93 , pp.742-755 (2010)
[Sakai 06]
酒井哲也: “よりよい検索システム実現のためにー正解の良し悪しを考慮した情報検索評価の動向ー”, 情報処理 Vol.47 No.2 (2006)
[Sakai 07]
酒井美里, “特許検索手法のマニュアル化と検索ノウハウの伝達”, 情報管理 50 , pp.569-577 (2007)
[Sasaki 04]
佐々木裕, 磯崎秀樹, 鈴木潤, 国領弘治, 平尾努, 賀沢秀人, 前田英作: “SVM を用いた学習型質問応答システム SAIQA-II (自然言語)”, 情報処理学会論文誌 45, pp.635-646 (2004)
[Sato 06]
佐藤祐介, 岩山真: “引用情報に基づく特許文献の重要度算出方式の検討(情報検索・情報解析)”, 情報処理学会研究報告. 情報学基礎研究会報告, 2006 ,pp.9-16 (2006)
[Sato 14]
佐藤祐介, 岩山真: “意図どおりの技術マップを作成するための特許自動分

類技術: 大量の特許を対話的に独自の観点で分ける支援技術（データによる分析と評価)", Japio year book, pp.212-215 (2014)

[Sebastiani 02]

F. Sebastiani: "Machine learning in automated text categorization", ACM computing surveys (CSUR) 34, pp.1-47 (2002)

[Shibata 02]

類語大辞典（講談社 2002 ）柴田 武 (編集), 山田 進 (編集)

[Shinmori 04]

新森昭宏, 奥村学, 丸川雄三, 岩山真: "手がかり句を用いた特許請求項の構造解析", 情報処理学会論文誌 45, pp. 891-905 (2004)

[Suzuki 13]

鈴木誠, 吉川大弘, 古橋武：”コレスポンデンス分析を用いた文書検索に関する検討", 日本感性工学会論文誌 12, pp.283-290（2013）

[Takagi 05]

高木徹, 藤井敦, 石川徹也: "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（情報検索)", 情報処理学会論文誌 46 , pp.1074-1081 (2005)

[Takemori 10]

竹森久美子: "特許情報を利用した技術動向分析技術に関する調査研究［課題研究報告書]", (2010)

[Takeuchi 02]

竹内孔一, 内山清子, 吉岡真治, 影浦峡, 小山照夫: "語彙概念構造を利用した複合名詞内の係り関係の解析（< 特集> システム LSI の設計技術と設計自動化)", 情報処理学会論文誌 43 ,pp.1446-1456 (2002)

[Tan 06]

P.N. Tan, M. Steinbach, V. Kumar:" Introduction to Data Mining", Pearson Addison Wesley Boston (2006)

[Tokkyocho 12]

特許庁, "工業所有権法< 産業財産権法> 逐条解説", 発明推進協会, (2012)

[Tokkyocho 14]

特許庁, “平成 25 年度我が国における技術革新の加速化に向けた産業財産権の出願行動等」に関する分析調査報告書”, 特許庁 (2014)

[Tomiura 96]

冨浦洋一, 日高達:"k-NN 推定法に基づく統語的曖昧さの解消法",電子情報通信学会技術研究報告(1996)

[Uchiyama 06a]

内山清子, 石崎俊: “特許文に含まれる複合名詞の解析”, 言語処理学会第 12 回年次大会発表論文集, pp.1107-1110 (2006)

[Uchiyama 06b]

内山清子, 石崎俊: “特許文に含まれる複合語の処理と検索への応用”, 特許情報活用の時代の検索と機械翻訳技術 Japio 年誌, 財団法人 日本特許情報機構, pp.90-93 (2006)

[Uchiyama 07]

内山清子, 栗飯原俊介, 石崎俊 : ”特許文書における複合語の意味関係解析”, Japio Year Book 2007,pp.160-165（2007）

[Uneda 12]

畝田将登, 難波英嗣, 竹澤寿幸, 乾孝司, 岩山真, 橋田浩一, 橋本泰一, 藤井敦:”特許請求項と詳細説明の自動対応付け”, 言語処理学会第 18 回年次大会（2012）

[Usui 13]

臼井裕一:” 日本における特許分類の問題点”, 情報の科学と技術 63(7), pp.1-6 (2013)

[Watanabe 03]

渡部勇: “テキストマイニングの技術と応用 (< 特集> 情報の分析・解析法)”, 情報の科学と技術 53, pp.28-33 (2003)

[Wipo 14]

World Intellectual property organization

http://www.wipo.int/portal/en/index.html (2015.03.15 参照)

[Wu 10]

C.-H. Wu, Y. Ken, T. Huang: “Patent classification system using a new

hybrid genetic algorithm support vector machine", Applied Soft Computing 10, pp.1164-1177 (2010)

[Yamada 96]

山田剛一, 森辰則, 中川裕志: "情報検索のための複合語マッチング", 情報処理学会研究報告. 情報学基礎研究会報告 96 ,pp.33-39 (1996)

[Yanagimoto 13]

柳本豪一：”ニューラルネットワークを用いた文書類似度の推定”, 第 27 回人工知能学会全国大会（2013）

[Yanaguchi 06]

日本語大シソーラス類語検索大辞典（大修館書店　2006）　山口翼

[Yamaoka 98]

山岡正輝, 岩城修, 馬場口登, 北橋忠宏: "構造の類似性に着目した対話型の非定型文書解析手法", 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解 98 ,pp.17-24 (1998)

[Yasuda 06]

保田明夫: "形態素解析と分かち書き処理",

http://wordminer. comquest. co. jp/wmtips/pdf/ (2006).

[Yoshida 05]

吉田耕一:”自動大分けシステムから中分けシステム 類似文献 探索を利用した自動分類付与ツール”, Japio 創立 20 周年記念誌, pp.86-89 (2005)

[Yoshifuji 97]

吉藤幸朔, "特許法概説（第 12 版）", 有斐閣 (1997)

[Yumoto 01]

湯本紘彰, 森辰則, 中川裕志: "出現頻度と連接頻度に基づく専門用語抽出", 第 145 回情報処理学会自然言語処理研究会資料, pp.111-118 (2001)

関連業績リスト

【学術論文】

［1］ 柳堀恭子, 津田和彦 : “拒絶査定情報を用いた類似特許検索知識の構築法”, 情報ディレクトリ学会誌, vol.13, 76-83 (2015)

［2］ Kyoko Yanagihori, Kazuhiko Tsuda : “Similar document determination method by using compound nouns ”, Asia Pacific Journal of Contemporary Education and Communication Technology (APJCECT), vol.1, 186-192(2015)

【国際会議】

［1］ Kyoko Yanagihori, Kazuhiko Tsuda : “Issues of the Morphological Analysis in Comparison with the Compound Noun Extraction Analysis for a Patent Document”, Information Systems International Conference, 2013ISICO 477-482 (ISBN : 978-979-18985-7-7), (2013.12)

［2］ Kyoko Yanagihori, Koji Tanaka and Kazuhiko Tsuda: “Improvement of Terminology Extraction Method for Specific Patent Search”, 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Procedia Computer Science 35 (2014) 879-885 (2014.9)

［3］ Kyoko Yanagihori, Kazuhiko Tsuda : “Verification of patent document similarity of using dictionary data extracted from notification of reasons for refusal”, COMPSAC 2015: The 39th Annual International Computers, Software & Applications Conference (2015.04 accepted)

【講演報告】

［1］ 柳堀恭子, 津田 和彦 : “特許拒絶理由通知書を利用した複合語類似の検証の一考察” , 第 62 回情報処理学会 電子化知的財産・社会基盤研究会,

Vol.2013-EIP-62 No.1 1-6 (2013.11)
［2］ 柳堀恭子，津田和彦：“特許文書中に出現する複合名詞の類似を判定する方法”，平成 26 年度電気学会電子・情報・システム部門大会,GS4-7 1527-1528 (2014.09)
［3］ 柳堀恭子，津田和彦：“特許情報から抽出した複合名詞を利用文書類似度の検証”，JSAI Special Interest Group on Business Informatics (SIG-BI) (2015.03)