特許中の語に着目した被引用数予測による重要な特許の発見

花谷　翔† 古屋　昭拓†† 大島　裕明††

† 兵庫県立大学社会情報科学部　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 兵庫県立大学大学院情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
E-mail: †{fa19j066,ad21m044}@guh.u-hyogo.ac.jp, ††ohshima@ai.u-hyogo.ac.jp

あらまし　本研究では，特許データを用いて重要な特許を発見する手法を提案する．具体的には，被引用数を予測することで重要な特許を発見する．被引用数を予測するために重要な特許に見られる 3 つの特徴を用いる．その特徴とは，汎用性があるという特徴，珍しい概念が含まれているという特徴，概念の珍しい組み合わせが含まれているという特徴である．汎用性がある特徴を発見するためには，引用元の分野を用いる．珍しい概念が含まれているという特徴，概念の珍しい組み合わせが含まれているという特徴を発見するためには，新たな技術用語や新たな技術用語の組み合わせが常に使われる一般的な語となるかに注目する．

キーワード　特許，テキストマイニング

1 はじめに

本研究で扱う特許という制度には，発明者や技術を保護する目的がある．特許とは特許法によって特許権をあたえることである．加えて，発明を公開する代わりにその発明を一定期間，独占的に使用することが出来る．科学分野では，日々，iPS 細胞，水素自動車や mRNA ワクチンのような新技術が生まれている．しかしながら，技術を保護しない場合，技術の模倣や，利益の横取りが起こりえる．それにより，モチベーションが低下し，産業発展の機会が減少してしまう．このような問題を解消するために，特許という制度が存在する．

特許には技術に関する情報が含まれるため，さまざまな分析を行うことが出来る．例えば，企業の経営方針，技術傾向の把握，自社で使える技術の発見などである．

しかし，特許の分析には時間とコストがかかる．なぜなら，特許は日本で年間約 30 万件，世界では年間約 300 万件も出願されているからである．その特許の中には永久機関の様な発明にならないものも含まれている．加えて，特許を分析するには技術に関する幅広い専門的な知識が必要である．そのため，専門的な知識のある者が時間を費やして特許データを分析しなければならない．

上記の問題を解決するため，重要な特許を発見する研究が行われてきた．特許の重要性を求める手法として被引用数が用いられる．被引用数とは基本的には他の特許から引用された数のことを指す．

しかしながら，特許は公開後すぐには被引用数が伸びにくい．そのため，公開後すぐの特許の評価に用いられにくいという問題がある．例えば，図 1 を用いて説明する．図 1 を作る際に用いた特許は 2002 年に公開された特許である．この特許が公開されてから 2 年経過したタイミングでは被引用数が 4 件ほどである．それ以降の被引用数の推移が図 1 の左のグラフの様に横ばいになっていくか，大きく伸びていくかはわからない．た

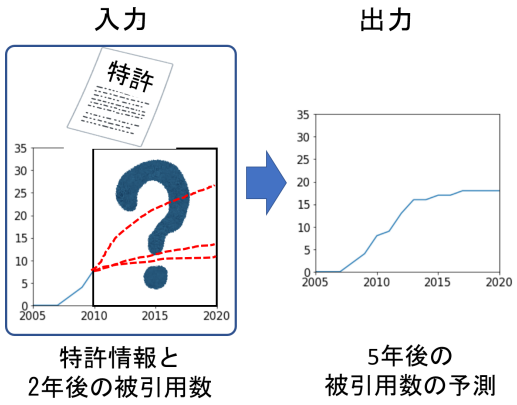

図 1　特許の情報と 2 年後の被引用数を用いて 5 年後の被引用数の予測

だし，ある程度年月が経過すると基本的には被引用数の上昇は止まり，特許の重要度がわかるようになる．そのため，既存研究では重要な特許はある程度年月が経過してから判定されてきた．しかし，数年も経つとより優れた技術が出てくるため，より早く重要な特許であるかを知りたい．そこで，数年後の被引用数を予測する手法を提案する．被引用数を予測することにより，早く特許が重要な特許であるかについて評価できると考えた．そして，さまざまな最適な条件が考えられるが，本研究では，現在より 2 年前に公開された特許の中で現段階で被引用数が 5 件の特許を対象とする。そして、対象の特許が 3 年後被引用数が 8 件を超えるか予測する問題に取り組んだ。

被引用数を予測するために本研究では特許の 3 つの特徴を用いる．本研究では以上の特徴が 1 つでも用いられている特許を被引用数が 8 件以上になる特許として予測している．

2 関連研究

2.1 文中の語に着目した研究

Qaiser らの研究 [6] では，TF-IDF を用いた単語と文書の関

連性を調べるための単語とドキュメントの関連性を調べる研究を行っている．確認され拡張されたテキストを基に，TF-IDF，テキストランク，LDA を組み合わせてキーワードを抽出する．融合型マルチタイプ特徴量組み合わせアルゴリズムの性能は，従来の単一アルゴリズムよりも優れている．

Qian らの研究 [7] では，BERT とマルチクラス特徴量融合に基づくキーワード抽出アルゴリズムを提案している．この研究では，BERT [2] モデルにより論文の様な背景資料からキーセンテンス集合を抽出する．次に，拡張されたテキストを基に，TF-IDF，テキストランク，LDA を組み合わせてキーワードを抽出する．融合型マルチタイプ特徴量組み合わせアルゴリズムの性能は，従来の単一アルゴリズムよりも優れている．

村田らの研究 [16] では，n-gram 解析を用いて，学生のレポートの類似度を求めた．実験としては，不正を行った学生のレポートと教科書やインターネットの web サイトを類似度測定行った．結果としては，高い類似性が見られていた．

山田らの研究 [10] では，Wikipedia や Web テキストから自動獲得した単語間の関係を用いることにより，2 つの文書間の類似性を評価した．結果としては，ランダムウォークによって暗黙的に単語の重要語が考慮できることを確認され，tf-idf などの指標を用いてエッジの重みづけをしなくてもよい．

木村らの研究 [17] では，入力テキストに出現する名詞間の共起と，各名詞句と入力テキストとの意味的類似度を基にエッジに重み付けグラフを構築し，グラフ内の名詞句に対し，TextRank を用いてキーフレーズらしさの値を算出した．そして，結果としては従来の共起グラフを用いた MultipartiteRank に比べ，僅かに高い精度での抽出が可能になった．

Omid らの研究 [8] では，意味的なテキストの類似度の自動測定というタスクを実行するためのさまざまなベクトル空間モデルの性能を評価した．そして問題としては，特許間の類似性をモデル化するというものに取り組んだ．そしてその際に TF-IDF，トピックモデル，ニューラルモデルを比較した．結果としては，単純な TF-IDF が正しい選択となり，LSI や D2V の様な，より複雑な埋め込み方法は，テキストが非常に凝縮され，類似性検出のタスクが比較的粗い場合のみ正しい選択となる．

2.2 特許マイニングに関する研究

野守らの研究 [13] では，テキストマイニングに PLSA とベイジアンネットワークという 2 つの人工知能技術を応用した新たなテキスト分析技術を特許データに応用している．

西山らの研究 [20] では，特許文章の中から新技術の可能性のある表現を用いているフレーズを自動的に抽出しており，その技術から商品が出てくるまでの期間や，その技術が与えるビジネス的な影響の大きさを推定している．

Fall らの研究 [3] では，自動分類システムの訓練とテストのための特許文書の新しい参照コレクションを確立している．これらのコレクションを確立する方法としては，さまざまな機械学習アルゴリズムを適用させている．

Lee らの研究 [5] では，特許を事前学習した BERT モデルをファインチューニングをすることで，特許分類を行っている．この手法では CNN と単語埋め込みを用いた手法を凌駕する精度を示している．また，特許の請求の範囲だけで分類タスクに十分であることが示された．

kang らの研究 [4] では，深層学習モデルを用いた特許先行技術検索を行っている．この研究では先行技術検索を行う際に，深層学習モデルの中でも BERT を用いたモデルを用いることで質の低い特許を除去することに成功している．

上村らの研究 [19] では，特許情報において，word cloud と共起ネットワークを同時に用いて，効率的に，広い分野で，素早く課題・解決手段を抽出できる手法の検討を行った．結果としては，word cloud で抽出した「〜性」から共起ネットワークを構築することにより，各課題をどのような手段で解決しているのかを俯瞰的に理解できた．

安藤らの研究 [15] では，テキストマイニングを行って，あらかじめ分類体系を決めて，決めた分類体系に自動分類する検討を行った．結果としては，多次元尺度法による非類似度を用いた文書間のクラスタリングの特性として予め決められた分類体系に当てはめる自動分類への応用には限界があるという考えに至った．

2.3 重要な特許に関する研究

後藤らの研究 [12] では，特許の被引用数や出願者名の数を用いることで，特許の重要度を効果的に図ることができている．しかし，被引用数に関しては経過した年度によっても変わることを示している．

佐藤らの研究 [14] では，被引用数は重要な特許を発見する際に有効であることを示している．それに加えて，引用の種別，後願審査結果や被引用された時期・期間等の他の観点と関連させることで質的にも評価することができることを示している．

佐藤らの研究 [18] では，重要度を図る際にも HITS アルゴリズムを用いることが最も精度が高いことを示している．また，特許固有の引用情報に関して興味深い結果を示している．それは，特許として認めるかを審査官が判断材料にした特許よりも出願者が特許を出願する際に用いた特許の方が重要特許の評価に適していることを明らかにした．そして，自社引用が重要特許を評価する上で重要であるということも示している．

2.4 将来の数値の予測に関する研究

Salvatore らの研究 [1] では，個別企業の将来の株価変動の大きさを予測する手法の提案を行った．この研究では，世界的な新聞社の記事から語彙を生成した．その後，ビジネスセクターにおいて特定の時間帯に市場に影響を与える単語の特定を行った．そこから企業に関連する統計的指標をの特徴を用いる．この特徴を用いる手法は既存のものより精度が良い．

Thomas らの研究 [9] では，論文の被引用数の予測を行っている．彼らの研究では最新の言語モデルと文脈に応じた論文の単語埋め込みをすることによって大規模の入力を扱うことことのできる SChuBERT を用いることによって最先端モデルよりも良い結果を得た．また，より良い結果を得るためには，入力

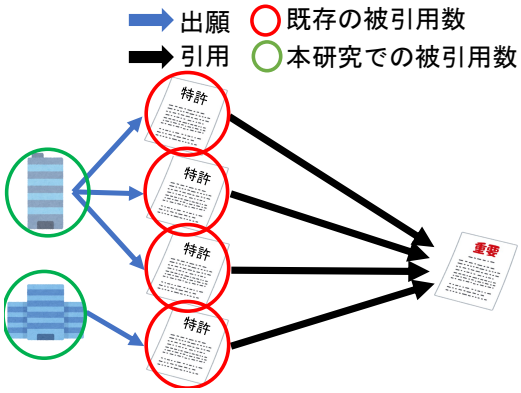

図 2　本研究が定義する被引用数の関係性について

長とデータ量の両方が重要であることも示されている．

それ以外にも関らの研究 [11] では，引用した論文の被引用数が多い論文は被引用数が伸びる可能性が大きいということを示している．このような特徴を用いてある一定の精度で被引用数の予測を行っている．

3　重要な特許の定義と問題定義

3.1　重要な特許の定義

重要な特許は多くの人に画期的な発明として認められる特許と一般的には言われている．しかし，どのように評価するかという問題がある．その問題を先行研究 [12] では被引用数を用いて判断していた．そのため，本研究では被引用数が多い特許を**重要な特許**であると定義する．しかし，被引用数のみを用いることにはさまざまな問題がある．まず，1 つ目は最新の特許を重要か推定できない問題である．この問題は第 1 章で述べたように被引用数を予測することによって解決する．

次に，2 つ目は，自社の特許を引用できるという問題である．企業によっては自社の特許を多く引用する．その結果，その企業のみが必要とする特殊な特許の被引用数が多くなる場合がある．そのため，被引用数の多さが多くの人に画期的な発明として認められている特許を示さない場合がある．

よって，本研究では既存の研究で用いられてきた被引用数ではなく，引用元特許を出願した団体数を，重要であるかの指標に用いる．例えば，図 2 のような関係図があったとする．この際，既存の重要な特許を発見する研究ではある特許を引用している特許が 4 つあることを用いて評価している．しかし，本研究ではある特許を引用した企業が 2 社存在することを用いて評価する．以後，この指標を**被引用数**と呼称する．

3.2　問 題 定 義

本研究では現在を y 年として扱い、$y-2$ 年に公開された特許を対象として扱う。対象の特許の中でも現在の段階で被引用数が i 件の特許を手法の評価の対象とする．そして、その対象の特許の被引用数が $y+5$ 年時点で，一定の閾値以上になるかの 2 値分類問題に取り組む。

3.2.1　被引用数が 8 件以上の特許を重要とする理由

本研究では重要な特許を判定する閾値を被引用数 8 件とする．

図 3　本研究の汎用性について

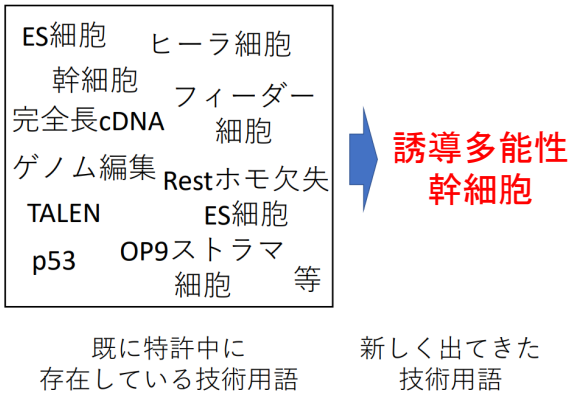

図 4　珍しい概念があるという特徴の説明

この被引用数 8 件以上という値は，全特許データの 0.1 パーセントほどであり，1 年間に出願されている特許数でいうと上位 300 件ほどに該当する．この上位 300 件に特許数を絞ることは第 1 節で述べたような，特許を分析する際の時間とコストがかかるという問題点の解決策として有効だと考えられる．

3.2.2　公開 $y+2$ 年時点で予測する理由

特許が公開されてから 2 年ほどでは，特許の被引用数の上昇が止まっていないことがある．しかし，公開 3 年を超え始めると，被引用数の上昇が横這いになる特許が増える．よって，本研究では公開 $y+2$ 年時点で予測を行う．

4　重要な特許の特徴

本研究では以下の 3 つの重要な特許の特徴に注目する．

- **汎用性があるという特徴**
- **珍しい概念が含まれているという特徴**
- **珍しい概念の組み合わせであるという特徴**

これらの重要な特許の特徴の説明をこの節では行う．

4.1　特許の汎用性

本研究でいう汎用性とはある特許を応用できる分野の幅のことを指す．例えば，図 3 の様な例が当てはまる．この図は，日立製作所が出願している特許第 3609590 という重要な特許を基にしている．この特許では,「動く物体の動作や行動を正しく認識させること」に関する発明であり，さまざまな分野で使われている．例えば，ウェアラブル端末，自走ロボット，車両用歩

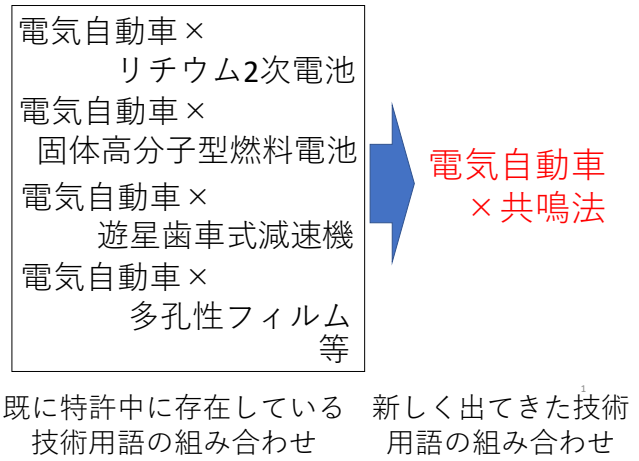

図 5　概念の珍しい組み合わせの説明について

行者検知システムなどが挙げられる．この例の様にさまざまな分野に使われる特許は重要になりやすいと考えた．

4.2　珍しい概念が含まれているという特徴の説明

この特徴は珍しい技術用語が含まれているという特徴である．例えば、この特徴は図 4 の様な例が当てはまる．この図 4 の例は，山中伸弥教授が出願した特許第 5098028 という特許を基にしている．この特許第 5098028 は誘導多能性幹細胞（iPS 細胞）の製造法の特許である．誘導多能性幹細胞は再生医療の分野において注目されている最新技術であるため，このような新たな技術用語を含む特許は重要であると考えられる．

本研究における珍しいの概念とは下記の 2 つの条件を満たすものである．

- **珍しい名詞句**
- **被引用特許に含まれている名詞句**

珍しい名詞句とは，今までの特許になかったような名詞句である．例えば，特許第 5098028 では誘導多能性幹細胞という名詞句が当てはまる．なぜなら，この特許公開時には誘導多能性幹細胞という名詞句はなかったからである．また，被引用特許に含まれている名詞句とは，引用された特許に珍しい名詞句が含まれているうことである．

誘導多能性幹細胞の様に，上記の 2 つの条件を満たす名詞句を珍しい名詞句とする．

4.3　概念の珍しい組み合わせであるという特徴の説明

珍しい概念の組み合わせであるという特徴は，図 5 の様な例が当てはまる．この例はトヨタ自動車株式会社が出願した特許第 4453741 という特許を基に作成している．この特許第 4453741 は電気自動車の充電に共鳴法を初めて用いた特許である．この特許は電気自動車の分野において注目されている技術の組み合わせを行った特許であるため，このような珍しい概念の組み合わせを含む特許は重要であると考えられる．

本研究における概念の珍しい組み合わせとは下記の 4 つの条件を満たすものである．

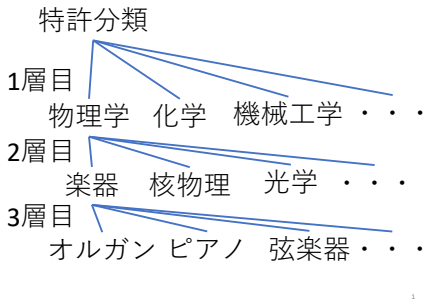

図 6　IPC（国際分類番号）の階層構造の一部

- **特許の構成上重要な名詞句の組み合わせ**
- **今までに使われてきた名詞句**
- **名詞句の珍しい組み合わせ**
- **被引用特許に使われる名詞句の組み合わせ**

特許の構成上重要な名詞句の組み合わせとは，特許の軸となる名詞句の組み合わせである．例えば，特許第 4453741 では共鳴法と電気自動車という名詞句の組み合わせである．

また，今までに使われてきた名詞句とは，今までの特許に存在する名詞句である．例えば，特許第 4453741 で出てきている共鳴法や電気自動車という概念も今までの特許中でよく使われている概念である．

珍しい名詞句の組み合わせとは，今までの特許にほとんど表れていない名詞句の組み合わせである．例えば，特許第 4453741 では共鳴法と電気自動車という名詞句の組み合わせは出願以前の特許に出てきていない．

被引用特許に使われる名詞句の組み合わせとは引用してきた特許に似たような概念の組み合わせが使われていることである．上記の 4 つの条件を満たす名詞句を初めての概念の組み合わせとする．

5　被引用数を予測する手法

本研究では，上記で述べたような特許の 3 つの特徴を用いて，被引用数を予測する手法を提案する．

まず，これらの特徴を発見する手法について以下の節で詳細を説明する．

5.1　汎用性を発見する手法

特許の汎用性を発見する手法について説明する．汎用性を発見するための提案手法としては，以下の様に行った．

（1）　**特許の分野の抽出**
（2）　**汎用性の計算**

5.1.1　特許の分野の抽出

特許の分野には引用元の特許国際特許分類を用いる．国際特許分類とは，特許文献の技術内容による分類を表すものである．この国際特許分類は階層構造である．例えば，図 6 の様な階層構造である．具体的には 1 層目では生活必需品，処理操作；運輸，化学；冶金，繊維；紙，固定構造物，機械工学；照明；加

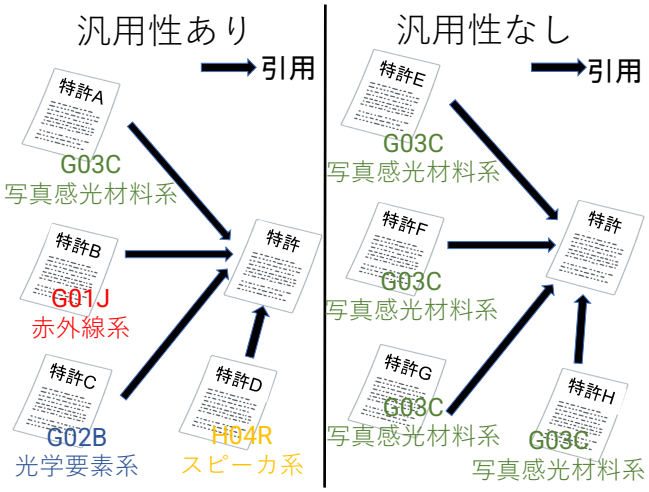

図 7　本研究の汎用性を求める手法について

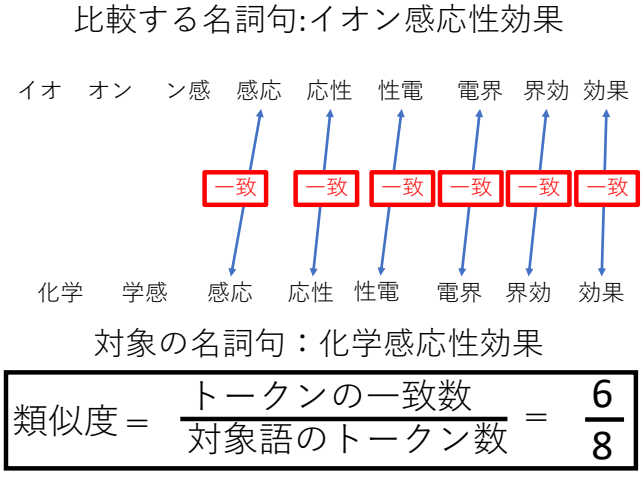

図 8　N-gram を用いた類似度の説明

熱；武器；爆破，物理学，電気の 8 つに分かれおり，その下に 5 層に細分化された分類が存在する．本研究では国際特許分類の階層構造の中の 3 層目までを特許の分野として用いる．図 6 で言うと，オルガン，ピアノ，弦楽器の層である．

5.1.2　汎用性の計算

本研究では，引用元の特許の分野の種類数によって汎用性があるかを判断する．汎用性があるか判断する閾値を公開後 2 年以内の被引用特許の 4 桁目の含めた国際特許分類の数が 4 種類以上，かつ公開後 2 年以内の被引用特許の 1 桁目の国際特許分類数が 2 種類以上とする．

例えば，図 7 のように，4 つの特許から引用されている特許がある．引用している特許にはそれぞれ国際分類番号が振られている．図 7 の場合，公開後 2 年以内の被引用特許の 4 桁目までを含めた国際特許分類数が 4 種類，かつ公開後 2 年以内の被引用特許の 1 桁目までを含めた国際特許分類数が 2 種類である．このような場合は汎用性がある特許とする．

5.2　特許中の名詞句の抽出

この節では，特許中の名詞句の抽出する方法について説明する．特許中の名詞句は，形態素解析の手法を用いて抽出する．形態素解析の手法の中でも，本研究では MeCab を用いる．MeCab とはオープンソースの形態素解析エンジンである．その MeCab で形態素解析した後,「名詞」,「接頭辞」,「接尾辞」からなる語を名詞句として抽出する．ただし，本研究では stopword リストを作成しており，そこに明記されている単語は抽出しないものとする．この stopword リストには,「前記」,「該」,「当該」,「上記」という名詞句や，方向を示す語，数値を示す語，場所を示す語，部位を示す語，状態を示す語で構成されるリストにおよそ 1000 語入っている．その中でも，表 1 は stopword の中でもよく使わていたものを挙げた．それに加え，本研究では特許を分析した結果 3 文字以下の名詞句には基本的に意味的に広すぎる語が多いので，抽出しない．

5.3　一般化した珍しい概念の発見手法

この節では，珍しい概念の発見する手法について述べていく．珍しい技術用語の発見するための提案手法としては，以下の順に行う．

（1）　**珍しい名詞句の抽出**
（2）　**珍しい名詞句の判定**
（3）　**珍しい名詞句が被引用特許中に存在するか**

．

5.3.1　珍しい名詞句の抽出

この節では珍しい名詞句の抽出について説明する．

まず，本研究では IDF(Inverse Document Frequency) を用いて珍しい名詞句の抽出を行う．IDF とは,「全文書中」で「ある単語を含む文書」が「どれくらい少ない頻度で存在するか」である．

そして，IDF は以下の様な集合から構成される．すべての特許 d から 5.2 節の手法を用いて抽出した名詞句を t とする．次に，1990 年から 2010 年 5 月に公開された特許の請求項 1 の集合を M とする．そして，本研究で用いる IDF 以下な計算式を用いる．

$$idf(t) = \log \frac{1+M}{1+df(t)} + 1$$

そして，以下の様な条件の名詞句 t を珍しい名詞句とする．

$$idf(t) >= 14$$

5.3.2　珍しい概念の判定

この節では珍しい概念の判定について説明を行う．

初めに，特許の請求項 1 中の「あって」,「おいて」という語に着目する．この請求項 1 中の記載で「〜において,」「〜であって,」という記載がある場合には，それ以前の部分にある名詞句をすべて珍しい概念であるか判定する名詞句の対象としない．

次に，本研究では N-gram を用いて図 8 の様に名詞句の類似度を測ることにより，珍しい概念の判定を行う．

対象とする特許 d から 5.3.1 節で抽出した珍しい名詞句を u とする．そして，特許集合 M の特許 m から 5.2 節で抽出した名詞句を p とする．そして，本研究では名詞句 u と名詞句 p の類似度を図 8 の様に計算する．そして，その類似度が 0.8 を

表 1　stopword リストの一部

請求項，記載，条件，以上，未満，下記，うち前記，値，方向，向こう，何人，同じ，感じ
手段，側，員環系内，基，式，最大，装着部，七角形，装着，側，方向，演出，部，型，機能，的，用，変化量，量，データ，中継機能
演出実行，決定手段，手段，粒子，報告，電極部，パラメータ，パレット，組成物，演出，線，要求，素子，決定，面上，用，前，材料
決定，度評価，度，評価，計測，情報，変化量，測定，樹脂，消費量，素子，樹脂，走行，中継，評価，寛骨，結露，対向標的，送達
変位，運動，成分，再始動，群，一方，各々，特徴，下記，方法，入力，出力，各，それぞれ，上，下，以外，内，以前，中，やすい
うち，とも，とき前記，後，以上，以下，毎，値，率，装置，面，実行，値，制御，状態，層，実行，数，間，範囲，系，ページ，パターン
矢印，末，点表示，判定，向き，自，工程，部分，部位，基盤，個，別，脚，手，足，係合，濃度，付近，部材，位置，温度，平均，媒体
軸，配列，位置，複数，本体全，こと，これ，それぞれ，ところ，もの，上，中，下，字，各，第，他，百，千，万，億，兆，下記，上記
時間，今回，前回，場合，一つ，年生，自分，ヶ所，ヵ所，カ所，箇所，ヶ月，ヵ月，カ月，箇月，名前，本当，確か，時点，全部，関係
近く，方法，我々，違い，多く，扱い，新た，その後，半ば結局，様々，以前，以後，以降，未満，以上，以下，幾つ，毎日，自体

超えた場合，名詞句 u と名詞句 p は似ている名詞句とする．これを特許集合 M の全特許に行い，似ている名詞句を持つ特許数が 2 つ以下の場合のみ珍しい概念と判定する．

5.3.3　珍しい概念が被引用数特許に存在するか判定

この節では珍しい概念が被引用特許中に存在するか判定する手法について説明する．

まず，本研究では N-gram を用いて図 8 の様に名詞句の類似度を測ることにより珍しい概念が被引用特許中に存在するか判定する手法を行う．

対象とする特許 d から 5.3.2 節で判定した珍しい概念を e とする．次に，1990 年から 2012 年 5 月に公開された特許の集合を N とする．この特許集合 N 中で特許 d を引用している特許を n とする．そして，特許 n から 5.3.1 節の様に抽出した名詞句を s とする．そして，本研究では概念 e と名詞句 s の類似度を図 8 の様に計算する．そして，その類似度が 0.8 を超えた場合，概念 e と名詞句 s は似ている名詞句とする．これを特許集合 N の特許 d を引用している全特許に行い，似ている名詞句を持つ被引用特許がある場合，概念 e を一般化した珍しい概念とする．

5.4　概念の珍しい組み合わせの発見手法

この節では概念の初めての組み合わせの発見手法について述べていく．初めての技術用語の発見するための提案手法としては，以下の順に行う．

（1）　頻出はしていないが，ある程度使われてきた名詞句の抽出

（2）　特許の構成上重要な名詞句の組み合わせの抽出

（3）　名詞句の珍しい組み合わせの抽出

（4）　概念の珍しい組み合わせが被引用数特許中に存在するか判定

5.4.1　頻出はしていないが，ある程度使われてきた名詞句の抽出

頻出はしていないが，ある程度使われてきた名詞句の抽出の方法について説明する．対象特許 d から 5.2 節の手法を用いて抽出した名詞句を t とする．そして，名詞句 t を 5.3.1 節と同じ条件で，IDF を計算した．そして，以下の様な条件の名詞句 t を今までに使われてきた名詞句とする．

$$idf(t) < 14$$

$$idf(t) >= 10$$

5.4.2　特許の構成上重要な名詞句の組み合わせの抽出

この節では，特許の構成上重要な名詞句の組み合わせの抽出を行う手法について説明する．

ここでいう，特許の構成上重要な名詞句とは，対象の特許では，出てくる名詞句ではあるが，この特許以外ではほとんど出てこない名詞句のことを示す．

このような構成上重要な名詞句を抽出する手法として，本研究では TF-IDF を用いる．TF-IDF では，対象特許 d としたとき， 5.2 節の手法を用いて抽出した名詞句を t に対応する次元の重みが以下の式で計算される．

$$tfidf(t,d) = tf(t,d) \times idf(t)$$

$$tf(t,d) = \frac{P(t,d)}{\sum_{i=1}^{n} P(t,d_i)}$$

$$idf(t) = \log \frac{1+M}{1+df(t)} + 1$$

そして，$tfidf(t,d)$ の値が上位 5 件の名詞句を抽出して作った名詞句集合を g とする．次に，作った名詞句集合 g を総当たり的に 2 つずつの名詞句の組み合わせを作る。それを名詞句の組み合わせ i とする．

5.4.3　名詞句の珍しい組み合わせの抽出

この節では名詞句の珍しい組み合わせの抽出について説明する．まず，本研究では IDF を用いて名詞句の珍しい組み合わせの抽出を行う．そして，IDF は以下の様な集合から構成される．対象特許 d から 5.4.2 節の手法を用いて抽出した名詞句の組み合わせを i とする．次に，1990 年から 2010 年 5 月に公開された特許の請求項 1 の集合を M とする．そして，この手法で用いる IDF の計算式は以下ある．

$$idf(i) = \log \frac{M}{df(i)}$$

そして，以下の様な条件の名詞句の組み合わせ i を珍しい名詞句の組み合わせ f とする．

$$idf(i) >= 14$$

5.4.4 名詞句の珍しい組み合わせが被引用数特許中に存在するか判定

この節では名詞句の珍しい組み合わせが被引用特許中に存在するか判定する手法について説明する．

まず，本研究では N-gram を用いて図 8 の様に名詞句の類似度を測ることにより珍しい概念が被引用特許中に存在するか判定する手法を行う．

対象とする特許 d から 5.4.3 節で判定した名詞句の珍しい組み合わせを f とする．次に，1990 年から 2012 年 5 月に公開された特許の請求項 1 の集合を N とする．この特許集合 N で特許 d を引用している特許を n とする．そして，特許 n から 5.3.1 節で抽出した名詞句を s とする．そして，本研究では名詞句の珍しい組み合わせ f と名詞句 s の類似度を図 8 の様に計算する．特許 n の中の名詞句で，名詞句の珍しい組み合わせ f のそれぞれの名詞句と類似度が 0.8 を超える場合，特許 n の中の名詞句の組み合わせと名詞句の珍しい組み合わせ f は似ている名詞句とする．これを特許集合 N の特許 d を引用している全特許に行い，名詞句の珍しい組み合わせ f を持つ被引用特許がある場合，名詞句の珍しい組み合わせ f を概念の珍しい組み合わせとする．

5.5 すべての特徴を含めた被引用数の予測手法

ここでは，被引用数が 8 件を超えるかを予測する方法について説明する．

- **汎用性があるという特徴**
- **珍しい概念が含まれているという特徴**
- **概念の珍しい組み合わせであるという特徴**

本研究では，上で示した，3 つの特徴のうち 1 つでも含む特許を被引用数が 8 件を超える特許と予測する．

6 実験評価

この節では，データセット，実験の設定，結果について述べていく．

6.1 データセット

本研究ではデータセットとして，特許庁が配布している特許情報バルクデータを用いる．その中でも，本研究では特許が登録された日が 2004 年から 2020 年の特許を用いる．特許データの構成としては，書誌情報，明細書，特許請求の範囲，請求の範囲，詳細な説明等で，構成されている．書誌情報には，特許番号，出願番号，発明の名称，参考文献，国際特許分類，日本の特許分類などが明記されている．

6.2 実験の設定

本研究の設定としては，図 9 のように 2012 年 5 月時点で 2010 年 2 月から 2010 年 5 月に公開された特許の中で被引用数が 5 件の特許を対象とする．この条件の特許数は 104 件である．そして，その対象の特許が 2015 年時点で，被引用数が 8 件以上になるかの分類問題に取り組む．この条件の被引用数が

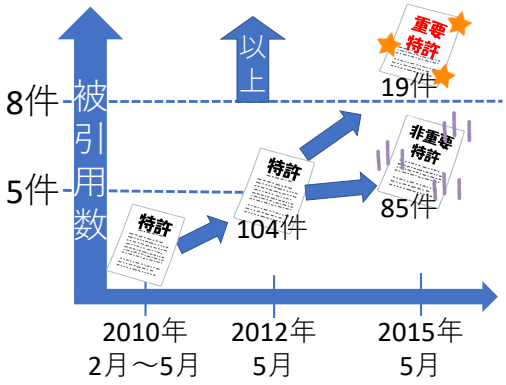

図 9 実験方法について

予測

正解		重要特許	非重要特許
	重要特許	11	8
	非重要特許	9	76

図 10 予測結果

8 件以上の特許数は 19 件である．

6.3 結　　果

それぞれの特徴を抽出した結果と予測した結果をここに記す．

6.3.1 汎用性を持つ特許

汎用性を持つ特許としては，バイオマスの加工方法の特許，ヘッドマウントディスプレイの特許，通信コイル構造の特許，大規模ＦＥＴアレイを用いた分析物測定の特許，リモートディスプレイ再現システムの特許，表示装置の特許，半導体装置の特許，通信システムの特許，電力需給システムの特許，面状光源装置の特許，照明装置の特許，アクセスゲートウェイ装置の特許，メディア情報の注目度測定装置の特許に汎用性が見られた．

6.3.2 一般化した珍しい概念

ここでは一般化した珍しい名詞句を発見する手法の結果を述べていく．対象の 104 件の特許から「化学感応性電界効果トランジスタ」「累計操作回数記憶」「炭化水素生産性微生物」「トラッキングパルス」の 4 つの名詞句が抽出された．

6.3.3 概念の珍しい組み合わせ

ここでは概念の珍しい組み合わせを発見する手法の結果を述べていく．対象の 104 件の特許から「幅射線照射，音波処理」，「幅射線照射，バイオマス原料」，「バイオマス原料，音波処理」，「オレフィン性二重結合，アクリル樹脂」，「オレフィン性二重結合，有機カチオン」，「リチウム二次電池負極，カーボンナノホーン」，「照明用レンズ，配光特性」，「カーナビゲーション，道路地図データ記憶」，「ウレタン樹脂，偏光子保護フィルム」の 9 つの概念の珍しい組み合わせが抽出された．

表 2　重要な特許と予測した結果

正解率	適合率	再現率	F 値
0.836	0.55	0.578	0.564

6.3.4 予測結果

3 つの特徴のうち 1 つ以上の特徴をもつ特許を正解と予測した結果が図 10 の様になった．この結果を正解率，適合率，再現率，F 値で計算すると表 2 の様になった．本研究の目的は，分析する特許数を少なくして，適合率を上げることである．その目的に対しては，ランダムに 20 個の特許を分析するよりはこの手法を用いて分析を行う方が重要な特許を分析する確率が高いという結果を得ることが出来た．

7 まとめと今後の課題

本研究では，被引用数を予測することで，重要な特許を発見する手法を提案した．そして，その手法では「汎用性」,「一般化した珍しい概念」,「概念の珍しい組み合わせ」という 3 つの特徴を 1 つでも持っている特許を重要な特許と予測する手法を提案した．その結果として，特許数を絞り込んで，適合率を上げるという目的に関しては，ある一定の精度があった．しかしながら，本研究の手法では文章の意味や，名詞句の意味を用いて，珍しい概念や概念の珍しい組み合わせを発見していない．それに加え，この手法が直近に出願された特許にも効果があるかも試していない．なので，今後の課題としては，word2vec や BERT の様な文章や名詞句の意味を理解することのできる深層学習モデルを用いて，珍しい概念や概念の珍しい組み合わせを発見していくことである．それに加えて，この手法が直近に出願された特許にも効果があるかに関しても行っていく必要がある．

謝　辞

本研究は JSPS 科研費 JP21H03775，JP18H03244，JP22H03905 の助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Salvatore M Carta, Sergio Consoli, Luca Piras, Alessandro Sebastian Podda, and Diego Reforgiato Recupero. Explainable machine learning exploiting news and domain-specific lexicon for stock market forecasting. *IEEE Access*, pp. 30193–30205, 2021.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[3] Caspar J. Fall, Atilla Törcsvári, Karim Benzineb, and Gabor Karetka. Automated categorization in the international patent classification. *Special Interest Group on Information Retrieval Forum*, pp. 10–25, 2003.

[4] Dylan Myungchul Kang, Charles Cheolgi Lee, Suan Lee, and Wookey Lee. Patent prior art search using deep learning language model. In *Proceedings of the 2020 Symposium on International Database Engineering & Applications*, pp. 1–5, 2020.

[5] Jieh-Sheng Lee and Jieh Hsiang. PatentBERT: Patent classification with fine-tuning a pre-trained BERT model. *arXiv preprint arXiv*, 2019.

[6] Shahzad Qaiser and Ramsha Ali. Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, pp. 25–29, 2018.

[7] Yili Qian, Chaochao Jia, and Yimei Liu. BERT-based text keyword extraction. In *Journal of Physics: Conference Series*, No. 042077, 2021.

[8] Omid Shahmirzadi, Adam Lugowski, and Kenneth Younge. Text similarity in vector space models: a comparative study. In *IEEE international conference on machine learning and applications*, pp. 659–666, 2019.

[9] Thomas van Dongen, Gideon Maillette de Buy Wenniger, and Lambert Schomaker. SchuBERT: Scholarly document chunks with BERT-encoding boost citation count prediction. *arXiv preprint arXiv:2012.11740*, pp. 148–157, 2020.

[10] 山田一郎, 宮崎勝, 住吉英樹, 古宮弘智, 田中英輝. ランダムウォークを利用した番組類似性評価. Technical Report 12, NHK 放送技術研究所, 2012.

[11] 関喜史, 松尾豊. 論文の引用情報を用いた論文被引用数予測. 人工知能学会 第 25 回全国大会 大会論文集, No. 1G22, 2011.

[12] 後藤晃, 玄場公規, 鈴木潤, 玉田俊平太. 重要特許の判別指標. RIETI ディスカッションペーパーシリーズ, 2006.

[13] 野守耕爾. テキストマイニングに複数の人工知能技術を応用した特許文書分析と技術戦略の検討. 情報の科学と技術, pp. 332–337, 2018.

[14] 佐藤貢司, 安井基陽, 田中厚子, 中村昭博, 中田守. 被引用情報を用いた重要特許抽出方法の検証: 時短でお手軽 特許分析! 情報プロフェッショナルシンポジウム予稿集 第 14 回情報プロフェッショナルシンポジウム, pp. 61–65, 2017.

[15] 安藤俊幸, 中西昌弘, 道中孝徳, 多田幸輔. アジア特許情報のテキストマイニングによる解析. 情報プロフェッショナルシンポジウム予稿集, pp. 13–17, 2011.

[16] 村田哲也, 黒岩丈介, 高橋勇, 白井治彦, 小高知宏, 小倉久和. 学生レポートの n-gram による類似度評価の検討. 情報科学技術フォーラム一般講演論文集, pp. 101–102, 2002.

[17] 木村優介, 楠和馬, 寺本優香, 波多野賢治. 単語埋め込みと名詞句の共起グラフを用いた教師なしキーフレーズ抽出手法の提案. Technical Report 2, 同志社大学大学院文化情報学研究科, 同志社大学大学院文化情報学研究科, 同志社大学文化遺産情報科学調査研究センター, 同志社大学文化情報学部, 2020.

[18] 佐藤祐介, 岩山真. 特許固有の引用情報を考慮した特許文献の重要度算出方式の検討. 情報管理, pp. 334–344, 2008.

[19] 上村侑太郎. テキストマイニングによる効率的な技術課題・解決手段の抽出手法の検討. 情報の科学と技術, pp. 29–33, 2022.

[20] 西山莉紗, 竹内広宜, 渡辺日出雄, 那須川哲哉, 前田潤治, 倉持俊之, 林口英治. 未来技術動向予測のための技術文書マイニング. 人工知能学会 第 21 回全国大会 大会論文集, No. 2H53, 2007.