特許マイニングによる類似度推定に基づく物質の新たな用途の発見

古屋　昭拓†　　山本　岳洋†　　窪内　将隆††　　大島　裕明†

† 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
†† 堺化学工業株式会社　〒 590–8502 大阪府堺市堺区戎島町 5–2
E-mail: †ad21m044@gsis.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp, †††kubouchi-m@sakai-chem.co.jp, ††††ohshima@ai.u-hyogo.ac.jp

あらまし　本稿では，ユーザが与えた物質名に対して，その物質名に関する新たな用途を発見する手法の提案を行う．基本的なアイデアは，機能が類似する物質は用途も類似するという考えに基づく．例えば，酸化チタンと酸化亜鉛という物質は顔料や充填剤，紫外線反射剤など共通する機能を多く持つため用途も類似している．このとき，酸化亜鉛のみが蛍光体として使われていれば，酸化チタンも同様に使える可能性がある．このような用途を新たな用途として発見する．提案手法では，2014 年までの特許を用いて新たな用途の発見を行う．2 つの実験を行った．類似する物質を特定できているか，用途を推定できているか確認した．化学分野の専門家によって類似する物質を特定できているか評価した．特許から新たに発見された用途を正解データとして用途の評価を行った．結果として，類似する物質を特定できており，新たな用途をわずかに発見できたが，今後，表記ゆれに対応することが課題であることがわかった．
キーワード　テキストマイニング，特許，情報推薦

1　はじめに

特許情報の利活用が進んでおり，実際に経営戦略に組み込まれた事例 [1] が多く存在する．特許情報を分析すれば，企業の強み，弱みを把握することや，新事業の着想を得ることができるため，多くの企業が取り組んでいる．

その中で，化学メーカーは物質の新たな用途を探すために特許情報を分析することがある．例えば，酸化チタンという物質の新たな用途を発見する場合，現在どのように使われているのか，今まではどうだったのかを大量の特許から把握する必要がある．しかし，特許は年間約 30 万件出願 [2] されており，正確かつ迅速に物質の用途を把握するのは困難である．また，様々な角度から分析を行い，新たな用途の候補を考える必要もあるが，これには専門的な知識を必要とし，特許を用いて物質の新たな用途を発見するには時間とコストがかかる．そのため，本研究では，特許などのビッグデータから物質の新たな用途を発見する手法の提案を行う．

本研究における理想的な物質の新たな用途とは，酸化チタンという物質に対して,「蛍光体として化粧品に使用する」という用途を発見することが望ましい．酸化チタンは酸化亜鉛という物質と機能が似ているため，塗料，化粧品など共通の用途で使わている．このとき，酸化亜鉛のみが蛍光体として化粧品に使われており，似た性質を持つ酸化チタンも同様に蛍光体として使える可能性がある．本研究では，このような用途を発見する手法を提案する．

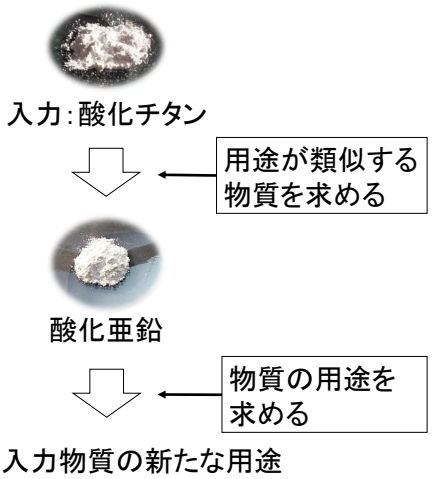

図 1　入力された物質の用途となる可能性があるものを推定する

本研究では,「物質名」を入力し，その新たな用途を発見する手法を提案する．新たな用途を発見する基本的なアイデアは,「類似する物質の用途は新たな用途となる可能性がある」という仮定に基づく．そのため，類似する物質の用途を求めることで，その中から新たな用途を発見することができると考えられる．概要を図 1 に示す．

本研究では，データセットとして特許庁が配布を行っている特許情報バルクデータ [3] を用いる．提供されている様々なデータの中でも，特許・実用新案公報情報（特許，実用新案登録）を

1：https://www.jpo.go.jp/support/chusho/document/bunseki/jirei2019.pdf
2：https://www.jpo.go.jp/resources/report/nenji/2021/document/index/honpenall.pdf
3：https://www.jpo.go.jp/system/laws/sesaku/data/

利用した．このデータは，特許庁によって登録された特許からなり，登録日が 2004 年から 2020 年の特許を対象としている．

実験として，類似する物質を特定できているか，新たな用途を推定できているか評価を行った．類似する物質は化学分野の専門家によって 10 種類のテストクエリを用いて評価を行った．新たな用途については，100 種類のテストクエリを用いて新たな用途を推定し，正解データによる評価を行った．データセット中の公開年が 2014 年までの特許を用いて物質の新たな用途を推定し，2014 年以降の公開年の特許から抽出される物質の用途を正解データとして手法の評価を行った．

2 関 連 研 究

本研究では，特許マイニングによって物質の新たな用途の発見を行う．技術を対象に新規用途の発見を行う研究 [13] [15] は多いが，物質を対象とする研究は少ない．また，類似物質の推定では，物質の分子構造の類似性を見る研究 [18] [10] [14] は多いが，自然言語の観点から類似性を推定する研究は少ない．特許文書内での語の共起や語の関係を表したグラフを用いて類似物質の推定を行うため，新規性があると考えられる．

2.1 特許マイニングに関する研究

村上の研究 [19] では，ワードクラウドや共起ネットワークといったテキストマイニング手法を用いることで，特許内で書かれている技術的課題や解決手段に関する記述を効果的に示している．Wang らの研究 [5] では，VSM，LSA といった検索手法を用いることによって，特許から新規性の検出を行っている．特許に対して外れ値検出を行うことで，新たな組み合わせを発見している．西山らの研究 [20] では，特許などの技術文書から当該技術によって実現できる事柄の抽出や提示を行っている．キーワードを用いて表現を抽出し，マップに表示することで，効果的に情報を提示している．

太田らの研究 [13] では，特許マイニングによって既存技術の新たな用途を発見する手法を提案している．発明が達成する「効果」と，解決する「課題」を抽出し,「課題」は類似するが「効果」は類似しないものを意外な「用途」の組み合わせとして出力している．文の類似度は word2vec を用いて求めており，新たな用途を発見することに成功している．野守の研究 [15] でも，既存技術の新たな用途を発見する手法が検討されている．特許内に出現する「用途」に関するトピックと「技術」に関するトピックの関係をベイジアンネットワークでモデル化することにより，関係性の強い「技術」を推定している．

近年，特許文書を対象とした，機械学習モデルを用いた研究も盛んに行われている．Fall らの研究 [2] では，機械学習のための特許分類タスク用データセットを作成している．そのデータセットを用いて，さまざまな機械学習アルゴリズムによる特許分類を行い，分類結果を示している．また，自然言語を対象としたタスクにおいて BERT [1] を用いた研究が増えており，特許を対象としたタスクにおいても，BERT を用いる研究がいくつも報告されている．Kang らの研究 [6] では，特許検索のタスクを行っており，この研究では，特許の先行技術調査の際，BERT を用いることでノイズとなる特許を除去することに成功している．Lee らの研究 [8] では，特許文書で追加学習した BERT を用いて，特許分類タスクを行っている．従来の深層学習モデルである CNN にワードエンベディングを組み合わせた手法と，特許文書で追加学習した BERT を用いた手法で特許分類の精度を比較している．その結果，BERT が特許分類タスクにおいて有効であることを示している．Lee らの研究 [7] では，自動生成された特許請求項の評価に BERT を用いており，GPT-2 によって生成された特許請求項を BERT を用いて評価し，特許請求項の自動生成タスクに取り組んでいる．

2.2 類似物質の推定を行う研究

高橋らの研究 [18] では，物質の性質の類似性を物質の構造全体の類似性から判断する手法を提案している．従来の手法では，物質の特定の部分構造の有無に注目するものが多く，事前に定義された部分構造に評価が大きく影響するため，構造全体に注目した手法の提案を行っている．Marwin らの研究 [10] では，LSTM に基づくリカレントニューラルネットワークを用いて，性質が類似する分子の生成を行っている．分子構造を SMILES 形式 [11] で表し文字列とすることで，統計的言語モデルで学習を行っている．和田らの研究 [14] では，化学物質の部分構造に注目することにより類似性を推定する手法を提案している．物質の各原子がどの部分構造に含まれるかを求め，その情報をもとに物質の構造の類似性を推定している．

2.3 類似度推定にグラフを用いた研究

人やモノの類似性を推定する手法として，グラフ構造を用いた研究が盛んに行われている．中山らの研究 [12] では，単語間の関係を表したグラフを用いて，内容が類似するテレビ番組を推定する手法を提案している．テレビ番組の番組概要文の類似性を単語間の関係を表したグラフを用いて推定している．文書間の類似性を推定する従来手法である Okapi-BM25 を用いた手法や tf-idf を用いた手法と比べ，提案手法は人手による結果に近いことが分かった．中辻らの研究 [17] では，ネットショッピングなどでドメインを跨るアイテムの推薦を行う手法の提案を行っている．同じアイテムを消費したユーザや SNS 上でやり取りのあるユーザの関係を表したグラフなどを用いて，Random Walk with Restart によってユーザがまだ消費していないドメインのアイテムとの関連性を推定している．

3 用語の定義

本節では新たな用途について述べる．本研究における用途について述べた後，新たな用途について説明する．

3.1 用　　途

本研究における**用途**は，

「機能名」として「カテゴリ名」に使用する

という形式で表すものとする．例えば，以下のようなものが挙げられる．

顔料として化粧品に使用する
融雪剤として街路の清掃に使用する
導電剤として電極に使用する

この形式で表すことにより，物質のどのような機能をどのようなカテゴリに対して使うのか表すことができ，物質の用途を適切に表せられると思われる．酸化チタンという物質を例に挙げて説明する．酸化チタンは，顔料として使われることが多く，一般的に酸化チタンの用途といえば顔料が挙げられる．しかし，酸化チタンは顔料として発色をよくする目的で化粧品や塗料，樹脂，繊維，食品など様々な対象で使用されており，単に顔料というだけでは何に対して使われているかわからない．また，酸化チタンは，化粧品に使われることも多く，これも酸化チタンの用途として挙げられる．しかし，化粧品は顔料や蛍光体，紫外線反射剤など様々な機能を持った物質が配合されていることがあり，単に化粧品というだけではどのような機能で使われているかわからない．そのため，用途は単に顔料や化粧品とするのではなく,「顔料として化粧品に使用する」といった「『機能名』として『カテゴリ名』に使用する」という形式で表すことが適切であると考えた．

3.2 用途の機能名

用途における**機能名**について述べる．本研究における機能名とは，物質の機能を表す語であり，例えば，顔料，紫外線反射剤，抗菌剤といった語が挙げられる．顔料であれば，物質が着色に適した機能を持つことを表し，紫外線反射剤であれば紫外線を反射する機能を持つことを表している．このような機能名は,「ある物質は○○として使用できる」に当てはめることができるような語とする．本研究では，このような物質と機能の関係を is-used-as 関係と呼称し，

物質名 is-used-as 機能名

上記の関係で表すことができる語句を機能名とする．例えば，酸化チタンという物質は紫外線反射剤や顔料としてよく使われている．日焼け止めには紫外線反射剤として使われているため，

酸化チタン is-used-as 紫外線反射剤

と表すことができる．また，塗料では顔料として使われているため，

酸化チタン is-used-as 顔料

と表すことができる．このように，紫外線反射剤と顔料は酸化チタンと is-used-as 関係が成り立つため，機能名であるといえる．一方，機能名に似た語に金属や酸化物という語がある．これらの語は酸化チタンと is-a 関係が成り立ち，

酸化チタン is-a 金属
酸化チタン is-a 酸化物

と表すことができる．これらの語の is-used-as 関係を考えてみた場合，酸化チタンは金属や酸化物と is-used-as 関係で表すことができないため，機能名ではないとする．

3.3 用途のカテゴリ名

用途における**カテゴリ名**について述べる．本研究におけるカテゴリ名は，商品カテゴリやサービスカテゴリを表す語を指す．具体的には，化粧品や医薬品，肥料，漂白，清掃などが挙げられる．本研究では，その定義域として，特許で使用されているテーマコード [4] を用いる．テーマコードは約 3,000 種類のテーマからなる．例えば，化粧料，医薬品製剤，植物の栽培など様々なテーマを持つ．テーマはある内容の特許をまとめたものであり,「化粧料」というテーマは,「ほおべに」,「おしろい」,「アイライナー」,「制汗剤」,「日焼け止め」など化粧品に関する特許をまとめたテーマである．日本の特許には，特許分類として File Index（FI）が付与されており，これは約 20 万種類に細分化された技術範囲を示している．テーマコードは．このような FI を技術的なまとまりごとに約 3,000 種類にまとめたものであるため，網羅的に定められていると考えれらる．

3.4 物　　質

本研究における**物質**は，新たな用途を求める対象となる化学物質や素材などを指す．具体的には,「酸化チタン」,「酸化亜鉛」といった化学物質や,「木粉」,「竹」,「みかん」といった素材など，様々な語が挙げられる．

3.5 新たな用途

以上をもとに，**新たな用途**について説明する．本研究における新たな用途とは，機能が類似する物質が持つ用途とする．例えば，酸化スズという物質は，酸化チタンと機能が類似しており，顔料，充填剤，半導体など共通する機能を多く持っている．このとき，酸化チタンは光触媒として外壁の塗装に使用されており，似た特徴をもつ酸化スズも同様に光触媒として外壁の塗装に使用できる可能性がある．実際，2016 年頃，林らの研究 [3] において酸化スズを用いた新たな光触媒が発見されている．本研究では，このような機能が類似する物質が持つ用途を新たな用途とする．以後，機能が類似する物質を**類似物質**と呼称する．

4 新たな用途の推定手法

本研究では，ある物質の新たな用途を類似物質の用途から発見する．そのため，共起やグラフを用いて類似物質を求め，類似物質の用途から新たな用途を発見する手法を提案する．

4.1 手法の概要

ある物質の類似物質の用途から新たな用途を発見する．まず，ある物質と各物質との類似度を計算することで類似物質を特定する．そして，各物質と各用途の適合度を計算することで各物質の用途を特定する．最後に，物質の類似度と用途の適合度によって各用途のスコアを求めランキングを作成する．このスコアは，高類似度の物質の高適合度の用途のスコアが高くなるように計算する．そうすることで，ランキングの上位にはある物質と関連性が高く，当たり前の用途が並び，順位が下がるにしたがって関連性はあるが見たことのない新たな用途が出現するランキングとなると考えられる．

前提として，物質とその用途がペアとなった情報を取得する

4：https://www.jpo.go.jp/system/patent/gaiyo/bunrui/fi/themecode.html

必要がある．そのため，本研究では特許文書を対象にその情報の抽出を行った後，その情報を用いて物質の類似度と用途の適合度を求めランキングを作成する．以下の手順で行う．

（1） **物質と用途の抽出**
（2） **物質の類似度の計算**
（3） **用途の適合度の計算**
（4） **ランキングの作成**

4.2 物質名と用途の抽出

特許文書から物質とその用途がペアとなった情報を抽出した．特許の明細書内の択一形式の文に注目することで物質名と機能名を抽出し，特許に付与されている FI からテーマコードを調べることでカテゴリ名を抽出した．

4.2.1 択一形式の文の収集

本研究では，物質名と用途を特許から抽出する必要がある．そのため，まず特許中の択一形式の文を収集した．特許では，マーカッシュ形式という択一形式の記述形式が使われることがある．これは，あるグループを設定し，そのグループ中から 1 つを選択できるようにする記述形式である．例えば,「顔料としては酸化チタン、酸化亜鉛、炭酸カルシウムが挙げられる」といった文がある．このような文では，酸化チタン，酸化亜鉛，炭酸カルシウムのいずれかから顔料を選択することができるように記述されている．特許では，この択一形式の文は「機能名としては物質名、物質名、...、物質名が挙げられる」という形式で書かれることがあるため，物質名と用途の抽出に適していると考えられる．本研究では，文中に「としては」を含み，その直前が機能名である文を択一形式の文とした．

以上を踏まえ，まず，特許の明細書中の文を対象に文中に「としては」というキーワードを含む文を収集した．次に,「としては」の直前の名詞句が機能名である文を択一形式の文として収集した．MeCab によって形態素解析を行い，名詞，接頭詞，記号からなる語句を名詞句として抽出した．その際，名詞句の頭に付く「前記，上記，該，うち，これら，当該，特定，その他，それら」の語を削除した．抽出した名詞句が機能名であるかを判断する際，一般化学物質・優先評価化学物質の用途分類の選択索引表の用語 [5] を利用した．以後，**用途分類用語**と呼称する．この用語は，経済産業省所管の製品評価技術基盤機構（NITE）によって作成されており，化学物質の機能や使われ方について産業界にアンケート等をした結果を踏まえて，代表的な名称が使われている．例えば，顔料，充填剤，触媒などの用語が挙げられる．これらの語は本研究における機能名である．そのため，用途分類用語を含む名詞句を機能名として抽出した．用途分類用語は 462 種類あり，特許から抽出される機能名としては，白色顔料，体質顔料，酸触媒などが想定される．

4.2.2 物質名と用途の抽出

収集した択一形式の文から物質名と用途を抽出した．物質名の抽出は，先行研究 [16] で用いた深層学習による抽出器を利用した．固有表現抽出（NER）タスクでファインチューニングした BERT モデルを使用した．

本研究における用途は，機能名とカテゴリ名の組み合わせによって表現することから，用途の抽出は機能名の抽出とカテゴリ名の抽出に分かれる．機能名の抽出は，4.2 節での択一形式の文の収集の際にすでに行っている．カテゴリ名の抽出は，択一形式の文が収集された特許の FI を利用した．特許に付与された FI の主分類からテーマコードを求め，そのテーマ名をカテゴリ名として抽出した．

4.3 物質の類似度の計算

本研究では，抽出した物質名と用途を用いてある物質に対する各物質の類似度を計算した．共起を用いた手法，物質名と機能名とカテゴリ名のグラフを用いた手法，物質名と用途のグラフを用いた手法，グラフエンベディングを用いた 2 種類の手法の合計 5 つの手法を提案する．

4.3.1 共起を用いた手法

共起を用いた手法では，物質名の共起を利用して類似度を計算する．本研究における共起とは，択一形式の文に共に出現したことを指す．択一形式の文で共起する場合，その物質名は共通の機能名を持っているため，共起するほど機能が類似する類似物質であると考えられる．

共起による類似度の計算には，期待相互情報量（EMI）を用いた．EMI が 0 以上の物質のみを類似物質とした．$P(x)$，$P(y)$ が物質名 x，y の抽出回数，$P(x,y)$ が x と y の共起回数とすると，EMI の計算式は以下の通りである．

$$EMI(x,y) = P(x,y)\log_2 \frac{P(x,y)}{P(x)P(y)}$$

4.3.2 物質名，機能名，カテゴリ名のグラフを用いた手法

特許から抽出した情報を利用し，グラフを作成することで類似度を計算する．物質名と機能名，機能名とカテゴリ名の間に枝のあるグラフを作成した．例を図 2 に示す．

このようなグラフにすることで，似ている機能をより多く持っている物質を適切に表現できると考えられる．機能には似ているものが存在している．例えば，顔料と着色剤などが挙げられる．これらはどちらも着色に関係する機能であるため，似ている機能であると捉えたい．そのため，物質名，機能名，カテゴリ名からなるグラフを作成した．機能名とカテゴリ名の間に枝があることで，顔料と着色剤であれば塗料や化粧料などのカテゴリ名と枝を持つため，関連性があることを表すことができると考えた．任意の物質名の類似物質を Biased PageRank を用いて求めた．

グラフは以下のノード集合から構成される．特許から抽出した物質名集合を M，機能名集合を F，カテゴリ名集合を C とする．また，物質名と機能名の間の枝については，物質名と共起した機能名すべてに対して相互に存在するものとし，機能名とカテゴリ名の間の枝については，用途として抽出されたすべてに対して相互に存在するものとする．物質名から機能名への枝集合を E_{mf}，機能名から物質名への枝集合を E_{fm}，機能名からカテゴリ名への枝集合を E_{fc}，カテゴリ名から機能名への枝集合を E_{cf} とする．

5：https://www.nite.go.jp/chem/risk/sakuin.pdf

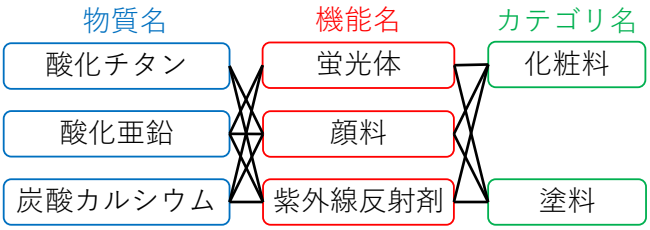

図 2　物質名と機能名とカテゴリ名からなるグラフにすることで物質と用途の関係を表すことができると考えられる

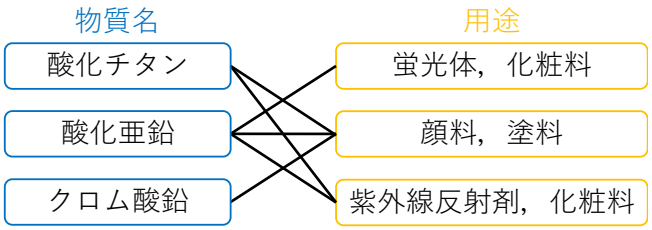

図 3　物質名と用途からなるグラフにすることでコンテキストを考慮することができると考えられる

このとき，グラフ $G = ((M \cup F \cup C), (E_{mf} \cup E_{fm} \cup E_{fc} \cup E_{cf}))$ を作成する．$m \in M$ の時，m からの枝を持つ機能名集合は $F_m = \{f | f \in F, (m, f) \in E_{mf}\}$ とでき，$F_m = \{f_1, f_2, f_3, \cdots, f_n\}$ とする．m から $f \in F_m$ への枝の重み $W(m, f)$ は，m と f の共起した回数に応じて決定され，m と f の共起回数を $P(m, f)$ とすると，

$$W(m, f) = \frac{P(m, f)}{\sum_{i=1}^{n} P(m, f_i)}$$

と計算される．他の枝の重みも同じように計算され，機能名とカテゴリ名の間の枝においては共起回数の代わりに用途として機能名とカテゴリ名が組み合わされた回数を使用している．

グラフ G に Biased PageRank を用いて類似度を計算した．t の時の各ノードのスコアを s_t，ジャンプ確率を α，エッジの重みの隣接行列を W，ジャンプ先を示す One-hot ベクトルを p とすると以下の式で表せる．

$$s_t = (\alpha - 1)(s_{t-1} W) + \alpha p$$

ジャンプ確率は 0.15 とし，ある物質名のみに 1 のスコアを与えた状態でスコアが収束するまで計算を行った．最終的に各物質名のノードが持つスコアを類似度とした．

4.3.3　物質名と用途のグラフを用いた手法

特許から抽出した情報を利用し，物質名と用途の二部グラフを作成することで類似度を計算する．例を図 3 に示す．

4.3.2 節のグラフでは，似ている機能を考慮するため，機能名とカテゴリ名の間に枝があった．しかし，顔料と染料のようにどちらも着色に関係する機能だが，性質が異なる機能である場合など，考慮出来ないケースが存在する．そのため，物質名と用途を直接的につなぐグラフを構築した．物質の類似度を SALSA [9] を用いて求める．

グラフは以下のノード集合から構成される．特許から抽出した物質名集合を M，用途集合を U とする．また．物質名と用途の間の枝については，物質名の用途すべてに対して相互に存在するものとし，枝集合を E とする．このとき，グラフ $G = ((M \cup U), E)$ を作成する．$m \in M$ の時，m からの枝を持つ用途は $U_m = \{u | u \in U, (m, u) \in E\}$ とでき，$U_m = \{u_1, u_2, u_3, \cdots, u_n\}$ とする．m から $u \in U_m$ への枝の重み $W(m, u)$ は，m の u の抽出回数である $tf(m, u)$ と，$idf(u)$ に応じて決定され，m と u の抽出回数を $P(m, u)$ とすると，

$$tf(m, u) = \frac{P(m, u)}{\sum_{i=1}^{n} P(m, u_i)}$$

$$idf(u) = \log \frac{|M|}{d^-(u)} + 1$$

$$W(m, u) = tf(m, u) idf(u)$$

と計算される．用途から物質名の枝の重みも同様に行い，重みはノードごとに合計すると 1 となるように正規化を行う．

グラフ G に SALSA を用いて任意の物質名の類似物質を計算した．t の時の各物質ノードのスコアを ms_t，各用途ノードのスコアを us_t，ジャンプ確率を α，ジャンプ先を示す One-hot ベクトルを p，用途ノードへのエッジの重みの隣接行列を W_{mu}，物質ノードへのエッジの重みの隣接行列を W_{um} とすると以下の式で表せる．

$$ms_t = (\alpha - 1)(us_t W_{um}) + \alpha p$$

$$us_t = ms_{t-1} W_{mu}$$

ジャンプ確率は 0.15 とし，ある物質名のみに 1 のスコアを与えた状態でスコアが収束するまで計算を行った．この手法では，物質名の評価を行うと同時に用途も評価しているため，任意の物質名に対して用途のランキングを得ることができる．そのため，以降のプロセスを省略してランキングを作成する．

4.3.4　グラフエンベディングを用いた手法

グラフエンベディングによって類似物質を計算する．各ノードのベクトル表現を DeepWalk を用いて求め，コサイン類似度によって任意の物質名との類似度を計算する． 4.3.2 節の物質名，機能名，カテゴリ名のグラフと 4.3.3 節の物質名と用途のグラフをそれぞれ用いた．そのため，グラフエンベディングを用いた手法を 2 種類提案する．各ウォークあたりのノード数は 10，ノード当たりのウォーク数は 80 とした．

4.4　用途の適合度の計算

類似物質の用途を特定するため，ある物質に対して各用途への適合度を計算した．適合度は，ある物質のメジャーな用途であれば高く，マイナーな用途であれば低くなるような値を想定している．本研究では，ある物質に対する用途の抽出回数を適合度として使用した．

4.5 ランキングの作成

本節では，ある物質名に対して物質の類似度と用途の適合度からランキングを作成する手法について述べる．物質の類似度と用途の適合度から高類似度の物質の高適合度の用途がランキング上位となるように用途にスコア付けを行う．まず，任意の物質名と各物質名との類似度を求める．類似度は合計が 1 となるように正規化を行う．この値を $ms_i(i = 1, 2, \cdots, n)$ とする．次に，各物質の用途の適合度を求める．各物質ごとに用途の適合度を合計 1 となるように正規化を行う．この値を $us_{ij}(i = 1, 2, \cdots, n)(j = 1, 2, \cdots, m)$ とする．物質の類似度と用途の適合度を用いて，各用途のスコアとして $s_j(j = 1, 2, \cdots, m)$ を計算する．

$$s_j = \sum_{i=1}^{n} ms_i * us_{ij}$$

スコアの高いものから用途のランキングを作成する．

5 実験と評価

5.1 データセット

本研究では，手法とその評価に特許庁が提供を行っている特許情報バルクデータ [6]を利用する．2004 年から 2020 年に登録された特許データのうち，公開日の情報を持つものを使用した．

この特許データ 1 件には，書誌情報，明細書，特許請求の範囲などの情報が XML 形式で格納されている [7]．書誌情報には，特許番号，出願番号，特許分類，発明の名称，公開日などが記されており，特許分類は，国際特許分類（International Patent Classification：IPC）と日本の特許分類（FI）に分けて記載されている．それぞれが，特許分類の主分類とそれ以外の分類に分かれて記載されている．主分類は特許に必ず付与されているが，それ以外の分類は付与されていないこともある．また，IPC はセクション，クラス，サブクラス，グループから構成される特許分類であり，FI は，IPC の構成要素に展開記号や分冊識別記号を加え，日本の特許のために IPC をより詳細にした特許分類である．明細書には，特許文書の本文が段落ごとに記載されており，特許請求の範囲には，請求項が段落ごとに記載されている．

5.2 評　　価

本研究では，新たな用途を発見する手法を 5 つ提案した．それぞれの手法について評価を行った．また，これらの手法は物質の類似度の計算手法がそれぞれ異なる．物質の類似度の計算手法の評価も行った．

5.2.1 物質の類似度の計算手法の評価

本研究では，物質の類似度の計算手法として 5 つの手法を提案した．それぞれの手法がうまく類似物質を特定できているか以下の手順で評価を行った．まず，各手法で類似度の高い上位 10 語をプールし，ランダムに並べる．これらの語に対し化学分野の専門家 1 名が類似物質かどうかのスコア付けを行った．スコアを付ける際は，類似物質ではない場合は 0，少し類似している物質の場合は 1，それなりに類似している物質の場合は 2，かなり類似している物質の場合は 3 とした．

評価には Normalized Discounted Cumulative Gain（nDCG）[4] を用いた．クエリとして 10 種類の物質名を用いた．酸化チタン，硫酸バリウム，酸化亜鉛，酸化マグネシウム，ビスフェノール A，チタン酸バリウム，二酸化ケイ素，グラフェン，窒化ガリウム，3 − メルカプトプロピオン酸の 10 種類である．

5.2.2 新たな用途を発見する手法の評価

本研究では，2014 年までに公開した特許文書から抽出した物質名やその用途の情報を活用することで，2015 年以降に公開された特許文書から抽出した物質の用途を推定可能か評価した．手法の出力は用途のランキングとなっているため PR 曲線と nDCG による評価を行った．

手法の評価には 100 種類の物質名をテストクエリとして使用した．Wikipedia の化合物一覧 [8]から抽出した物質名と NITE が配布している政府による GHS 分類結果 [9] の物質名称から抽出した物質名を対象とした．Wikipedia の化合物一覧からは 673 語，政府による GHS 分類結果の物質名称からは 3359 語を抽出した．物質名の末尾に付く価数は削除している．また，政府による GHS 分類結果の物質名称では，隅付き括弧とその中に記載された語句や，別名とその語句，再分類の文字列は名称から削除した．加えて，混合物という文字列を含む名称や塩は対象外としている．このようにして収集した物質名の全特許データの明細書内での出現回数を調べ，上位 5%の語を削除したランキングの上位から 100 語ごとに 10 区間に区切り，各区間からランダムに取り出した語をテストクエリとした．テストクエリを表 1 に示す．テストクエリの中で，特許から抽出されなかった物質と正解データがない物質がそれぞれ 6 種類あった．そのため，84 種類のテストクエリを用いて評価を行った．

正解データを 2 種類用意し，それぞれを用いて評価を行った．まず，2015 年以降に公開された特許から抽出される全ての用途を正解データとした．今回提案した手法では，出力のランキングの上位には入力物質の用途として当たり前の用途が出現し，ランキング下位になるにしたがって新たな用途が出現することが想定される．そのため，入力物質の当たり前の用途が出力されていることを確かめるために全ての用途を正解データとした．以後，この正解データを **一般用途正解データ**と呼称する．そして，2 種類目の正解データとしては，2015 年以降に公開された特許のみから抽出される新たな用途を正解データとした．以後，この正解データを**新たな用途正解データ**と呼称する．

PR 曲線を用いて 5 種類の手法を評価した．一般用途正解データを用いた．本研究の場合，recall が 1 とならないことがある．2014 年までに公開された特許から抽出される用途を用いてランキングを作成するため，2015 年以降に初めて出現す

6：https://www.jpo.go.jp/system/laws/sesaku/data/

7：https://www.jpo.go.jp/system/laws/koho/shiyo/document/kouhou_siyou_vol4-4/1-02.pdf

8：https://ja.wikipedia.org/wiki/化合物一覧

9：https://www.nite.go.jp/chem/ghs/ghs_download.html

表 1　新たな用途を発見する手法の評価のテストクエリ

プロピレングリコールモノメチルエーテル，トリフルオロ酢酸，チロシン，酸化鉄，アクリル酸メチル，ホルムアルデヒド，アンチモン
インターフェロン，メタクリル酸メチル，炭酸水素ナトリウム，メタクリルアミド，酪酸，ホルムアミド，硫酸カルシウム，インターロイキン
酸化銅，酸化クロム，塩化マグネシウム，アニソール，シトシン，チミン，ケモカイン，イソブテン，ヘキサン酸，ジペンタエリスリトール
ボラン，チオ硫酸ナトリウム，水酸化アンモニウム，イブプロフェン，シクロヘキシルアミン，トリプロピルアミン，クメン，ニトロメタン
アジ化ナトリウム，アスパルテーム，ナトリウムエトキシド，二酸化窒素，ペンタエリスリトールテトラアクリレート，リボフラビン
アセトアミノフェン，ジイソブチルケトン，ドデシルベンゼンスルホン酸，カプサイシン，ゲラニオール，炭酸亜鉛，硫酸カリウム，四塩化ジルコニウム
エチレングリコールモノエチルエーテルアセテート，ホウ砂，硫酸マンガン，トリエチレングリコールモノメチルエーテル，水酸化セシウム
ボルネオール，塩化マンガン，テアニン，五酸化バナジウム，コルチゾール，カルボン，メタクリル酸シクロヘキシル，テトラクロロエチレン
硫酸リチウム，フタル酸ジメチル，シトロネラール，フラボン，ジニトロベンゼン，アクリル酸イソオクチル，シンナムアルデヒド
ヒドラジン一水和物，エライジン酸，ジアリルアミン，アマンタジン，ヨウ化銀，オキシテトラサイクリン，過塩素酸ナトリウム，亜塩素酸
過マンガン酸，ニトロエタン，エフェドリン，亜リン酸トリエチル，ギ酸ナトリウム，塩化白金，シマジン，フルジオキソニル，ぎ酸
チロキシン，ブロモ酢酸エチル，フェロシリコン，メチラール，セレン化水素，次亜塩素酸カルシウム，シキミ酸，ラノステロール
過塩素酸アンモニウム，セレン酸，ホスチアゼート，クロロメタン，亜硝酸カリウム，チオセミカルバジド，臭素酸ナトリウム，フタル酸ジイソデシル

表 2　物質の類似度の計算手法の nDCG の結果

	nDCG@10
共起を用いた手法	0.744
物質名，機能名，カテゴリ名のグラフを用いた手法	0.629
物質名と用途のグラフを用いた手法	0.677
物質名，機能名，カテゴリ名のグラフにグラフエンベディングを用いた手法	0.801
物質名と用途のグラフにグラフエンベディングを用いた手法	**0.816**

る用途に対応できないためである．そのため，正解データのみに含まれる用途をランキングの末尾に追加することで recall が 1 となるようにした．各手法を用いて PR 曲線を求め，補完適合率を計算し，11 点補完平均適合率を用いて評価した．

nDCG を用いて 5 つの手法を評価した．2 種類の正解データを用いた．出力の適合度は，正解データに含まれる用途を 1，それ以外の用途を 0 とした．nDCG@200，nDCG@400，nDCG@600 をそれぞれ求めた．

5.3　結果および考察

物質の類似度の計算手法の評価結果を表 2 に示す．全ての手法である程度類似物質を推定できていることが分かった．物質名と用途のグラフにグラフエンベディングを用いた手法が最も良い結果となった．

新たな用途を発見する手法の評価結果を示す．一般用途正解データを用いた PR 曲線と nDCG による結果を図 4，表 3 に示す．PR 曲線による評価では，グラフエンベディングを用いた手法以外はある程度用途を発見できていることが分かる．物質名と機能名とカテゴリ名のグラフを用いた手法が最も良い結果となったが，共起を用いた手法や物質名と用途のグラフを用いた手法と大きな差はない．また，nDCG による評価においても同じような結果が得られた．そのため，物質名と機能名とカテゴリ名のグラフを用いた手法は他の手法と比べ，一般的な用途をランキング上位にうまく推薦していることが分かる．グラフエンベディングを用いた手法は，物質の類似度の計算手法の評価においては最も良い結果となったが，用途を推定する際には極端に悪い結果となった．原因としては，他の手法と比べて物質間の区別があまりうまくいっていないことが考えられる．グラフエンベディングでは酸化チタンの類似物質を求めた際，酸化チタンとの類似度が 0.9 以上の物質が複数見られた．類似度のランキング順に類似物質を見たところ，感覚的にはよいランキングが得られているものの，物質間の区別があまりうまく行えていない可能性がある．また，物質の類似度の計算手法の評価は上位 10 語を対象に行われているため，10 語以降は類似物質の推定がうまくできていないことも考えられる．

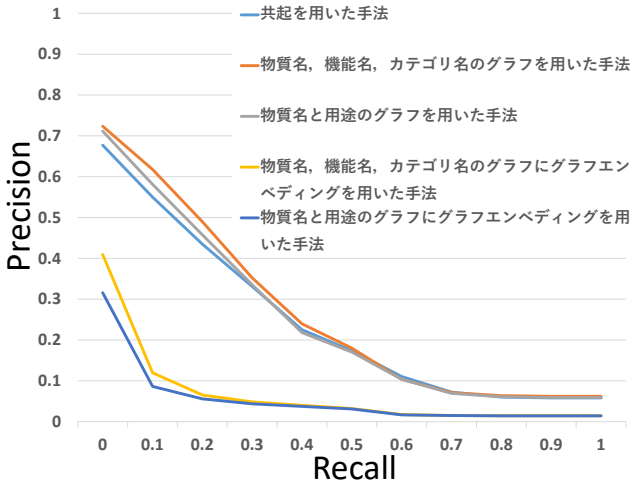

図 4　一般用途正解データを用いた 5 つの手法の PR 曲線の結果

新たな用途正解データを用いた nDCG による結果を表 4 に示す．共起を用いた手法が最も良い結果となった．共起を用いた手法では，特許の択一形式の文内で物質名が共起した回数をもとに類似物質を求めている．そのため，単に物質の機能的な類似のみを考慮することで類似物質を求めているといえる．他の手法と比べ，カテゴリ名については考慮しておらず，この違いが新たな用途の推定に適していたと考えられる．実際の出力において，酸化鉄が負極活物質として利用できる可能があることやアンチモンが半導体に利用できる可能性があることを確認できた．これらの用途は現在研究が盛んに行われている分野である．そのため，新たな用途の手がかりとなる出力を得ることに成功した．しかし，顔料と黒色顔料など，表現は異なるがほとんど同じ意味を表す語の区別ができておらず，うまく新たな用途が推定できていないものが大半をしめていた．今後，このような語の表記ゆれへの対応を行うことで，より明確に新たな用途のみを取り出すことができると考えられる．

表 3　一般用途正解データを用いた 5 つの手法の nDCG の結果

	nDCG@200	nDCG@400	nDCG@600
共起を用いた手法	0.414	0.438	0.449
物質名，機能名，カテゴリ名のグラフを用いた手法	**0.428**	**0.446**	**0.457**
物質名と用途のグラフを用いた手法	0.418	0.438	0.450
物質名，機能名，カテゴリ名のグラフにグラフエンベディングを用いた手法	0.152	0.176	0.190
物質名と用途のグラフにグラフエンベディングを用いた手法	0.132	0.154	0.170

表 4　新たな用途正解データを用いた 5 つの手法の nDCG の結果

	nDCG@200	nDCG@400	nDCG@600
共起を用いた手法	**0.133**	**0.153**	**0.163**
物質名と機能名とカテゴリ名のグラフを用いた手法	0.100	0.118	0.131
物質名と用途のグラフを用いた手法	0.111	0.130	0.144
物質名，機能名，カテゴリ名のグラフにグラフエンベディングを用いた手法	0.047	0.064	0.072
物質名と用途のグラフにグラフエンベディングを用いた手法	0.047	0.056	0.066

6 ま と め

本研究では，新たな用途を発見する手法の提案を行った．ある物質の類似物質の用途は新たな用途となる可能性があるという考えに基づき，特許から物質名と用途を抽出し，その情報から物質の類似度と用途の適合度を求めることで新たな用途を発見する手法を提案した．物質名の共起を用いた手法，物質名と機能名とカテゴリ名のグラフを用いた手法，物質名と用途のグラフを用いた手法，グラフエンベディングを用いた 2 種類の手法の 5 つの手法を提案した．その結果として，グラフエンベディングを用いた手法以外で，一般的な用途の発見がある程度できていることを確認した．また，新たな用途をわずかに発見することができた．今後，語の表記ゆれに対応することでより明確に新たな用途を取り出すことができると考えられる．

謝　　辞

本研究は JSPS 科研費 JP21H03775，JP21H03774，JP21H03554 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] 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.

[2] Caspar J. Fall, Atilla Törcsvári, Karim Benzineb, and Gabor Karetka. Automated categorization in the international patent classification. *SIGIR Forum*, Vol. 37, No. 1, pp. 10–25, 2003.

[3] Hayashi Hiroyuki, Katayama Shota, Komura Takahiro, Hinuma Yoyo, Yokoyama Tomoyasu, Mibu Ko, Oba Fumiyasu, and Tanaka Isao. Discovery of a novel Sn(II)-based oxide β-SnMoO4 for daylight-driven photocatalysis. *Advanced Science*, Vol. 4, No. 1, 2017.

[4] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, No. 4, pp. 422–446, 2002.

[5] Wang Juite and Chen Yi-Jing. A novelty detection patent mining approach for analyzing technological opportunities. *Advanced Engineering Informatics*, Vol. 42, Article 100941, pp. 1–11, 2019.

[6] 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, Article No. 1, 2020.

[7] Jieh-Sheng Lee and Jieh Hsiang. Measuring patent claim generation by span relevancy. *Computing Research Repository*, 2019.

[8] Jieh-Sheng Lee and Jieh Hsiang. Patent classification by fine-tuning BERT language model. *World Patent Information*, Vol. 61, Article 101965, pp. 1–4, 2020.

[9] Ronny Lempel and Shlomo Moran. SALSA: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, Vol. 19, No. 2, pp. 131–160, 2001.

[10] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, Vol. 4, No. 1, pp. 120–131, 2018.

[11] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, Vol. 28, No. 1, pp. 31–36, 1988.

[12] 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.

[13] 太田貴久, 南拓也, 山崎祐介, 奥野好成, 田辺千夏, 酒井浩之, 坂地泰紀. 特許文書を対象とした因果関係抽出に基づく発明の新規用途探索. 人工知能学会第 35 回全国大会, 2018.

[14] 和田貴久, 大野博之, 稲積宏誠. 部分構造に基づく構造類似性を用いた特徴抽出システムとその応用. 日本データベース学会論文誌, Vol. 7, No. 1, pp. 187–192, 2008.

[15] 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, Vol. 68, No. 7, pp. 332–337, 2018.

[16] 古屋昭拓, 山本岳洋, 窪内将隆, 大島裕明. 特許文書を用いた物質と特徴の関係理解に基づく物質の意外な用途の発見. 日本データベース学会第 14 回データ工学と情報マネジメントに関するフォーラム, 2022.

[17] 中辻真, 藤原靖宏, 内山俊郎. ユーザグラフ上のランダムウォークに基づくクロスドメイン推薦. 人工知能学会論文誌, Vol. 27, No. 5, pp. 296–307, 2012.

[18] 高橋由雅, 藤島悟志, 加藤博明. 化学物質の構造類似性にもとづくデータマイニング. 日本コンピュータ化学会, Vol. 2, No. 4, pp. 119–126, 2003.

[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, Vol. 72, No. 1, 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, 2007.